Light, Matter, and Geometry

The Cornerstones of Appearance Modelling

Frisvad, Jeppe Revall

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Light, Matter, and Geometry

The Cornerstones of Appearance Modelling

Jeppe Eliot Revall Frisvad

Kongens Lyngby 2008
IMM-PHD-2008-188

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Abstract

This thesis is about physically-based modelling of the appearance of materials. When a material is graphically rendered, its appearance is computed by considering the interaction of light and matter at a macroscopic level. In particular, the shape and the macroscopic optical properties of the material determine *how* it will interact with incident illumination. In this thesis the macroscopic optical properties are connected to the microscopic physical theories of light and matter. This enables *prediction* of the macroscopic optical properties of materials, and, consequently, also prediction of appearance based on the contents and the physical conditions of the materials.

Physically-based appearance models have many potential input and output parameters. There are many choices that must be made: How many material components to include in the model, how many physical conditions to take into account, whether the shape of the material should be coupled to the appearance model or not, etc. A generalised concept of shape and geometry is presented to provide a framework for handling these many degrees of freedom. Constraints between input and output parameters are modelled as multidimensional shapes. This gives the opportunity to use the appearance models not only for prediction, but also for *analysis* of the contents and the physical conditions of a material given information about its macroscopic optical properties. Since it is possible to measure these properties using camera technology, the presented framework enables analysis of material contents and conditions using camera technology.

Three detailed appearance models are presented as to exemplify the applicability of the theory: (1) A model which finds the appearance of water given temperature, salinity, and mineral and algal contents of the water; (2) a model which finds the appearance of ice given temperature, salinity, density, and mineral and algal contents of the ice; and (3) a model which finds the appearance of milk given fat and protein contents of the milk.

Resumé

Denne afhandling omhandler fysisk baseret modellering af materialers udseende. Når et materiale bliver fremstillet grafisk, beregnes dets udseende ved at overveje vekselvirkningen mellem lys og substans på et makroskopisk niveau. Mere præcist bestemmer materialets form og dets makroskopiske optiske egenskaber, *hvordan* det vil vekselvirke med indfaldende belysning. I denne afhandling kobles de makroskopiske optiske egenskaber til de mikroskopiske fysiske teorier for lys og substans. Dette muliggører *forudsigelse* af materialers makroskopiske optiske egenskaber, og derfor også forudsigelse af deres udseende v.h.a. materialernes indhold og deres fysiske tilstande.

Fysisk baserede modeller for udseende har mange potentielle input og output parametre. Der er mange valg, som må gøres: Hvor mange materialekomponenter der skal inkluderes i modellen, hvor mange fysiske tilstande der skal tages højde for, om materialets form skal koples til modellen for udseendet eller ej, o.s.v. Et genereliseret koncept m.h.t. form og geometri præsenteres for at give et system til behandling af disse mange frihedsgrader. Bindinger mellem et materiales input og output parametre modelleres som en multidimensionel form. Det giver mulighed for at bruge modeller for udseende ikke kun til forudsigelse, men også til *analyse* af et materiales indhold og dets fysiske tilstande givet information om dets makroskopiske optiske egenskaber. Da det er muligt at måle disse egenskaber ved brug af kamerateknologi, muliggører det præsenterede system også analyse af materialeindhold og tilstande ved brug af kamerateknologi.

Der præsenteres tre detaljerede modeller for udseende for at give eksempler på teoriens anvendelsesmuligheder: (1) En model til at finde vands udseende givet vandets temperatur, dets saltindhold og dets indhold af mineraler og alger; (2) en model til at finde is' udseende givet isens temperatur, saltindhold, massefylde og dens indhold af mineraler og alger; og (3) en model til at finde mælks udseende givet mælkens protein- og fedtindhold.

Preface

Study without thinking, and you are blind; think without studying, and you are in danger.

Confusius (552 B.C. – 479 B.C.), from the *Analects* (2:16)

This thesis was written out of curiosity and fascination with nature. It was prepared at the department of Informatics and Mathematical Modelling of the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in mathematical modelling.

The subject of the thesis is appearance modelling which is a subject within the branch of computer graphics known as realistic image synthesis. The student of realistic image synthesis is privileged in being allowed to investigate the reasons for all visual aspects of nature. Indeed all visual aspects are interesting, and the main objective is to capture their appearance correctly. It seems that this objective is most sensibly attained by physical models. However, the gap between the physicist's understanding of nature at the microscopic level of quantum particles and the modelling of appearance is wide. It is the aim of this thesis to build a bridge over the gap, or perhaps just to lay the foundation for the bridge. This is done in three parts: One part concerning light, one concerning matter, and one concerning geometry. A part on each of the three cornerstones of appearance modelling. Finally, there is a fourth part in which appearance models are developed based on the theory provided in the first three parts.

It is unusual for a thesis to spread over such a large variety of theories as you will find in this one. The usual approach would be to focus as narrowly as possible on the perfection of a single technique. The reason for the unusual approach is the gap and the missing bridge. If there had been a book closing the gap, or providing the bridge, this thesis would have been entirely different. As it is,

there is no such book. So instead of building a house with no foundation, this thesis became an attempt on laying the foundation itself.

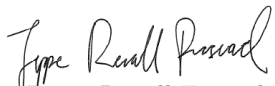
In graphics we would like to be able to model the appearance of as many different materials as possible. Therefore the theory has been kept as general and as flexible as possible throughout the thesis. It is in the effort to do so that the main contributions of the thesis appear. This makes it difficult to split the thesis in a part on background theory and a part on contributions. Instead, the introduction contains a relatively detailed overview of the thesis in which the contributions are pointed out.

The reader is assumed to have some graduate-level mathematical understanding and some general knowledge about graphics, in particular ray tracing.

The project was advised by Associate Professor Niels Jørgen Christensen and Professor Peter Falster who are both with the department of Informatics and Mathematical Modelling of the Technical University of Denmark. The project has been funded by a Ph.D. scholarship from the Technical University of Denmark. Part of the work presented in this thesis has been published, or is to appear. Details on these publications are provided in the Acknowledgements.

The human mind is a fantastic image processor. When we observe nature, we are able to draw conclusions based on subtle details. Especially if we know what to look for. Mathematical models which describe the appearance of nature have the ability to teach us what to look for. They can tell us the visual consequence of changing the contents of a material, or changing the temperature, or changing other properties. In this way, appearance models make observation of nature more instructive, but also more spectacular. Let us build the bridge and construct more models.

Lundtofte, November 2007



Jeppe Revall Frisvad

Acknowledgements

First and foremost thanks to Monica for her love and support, and for her coping with my absence at many long late hours. Also to my parents, Sven and Felicia, for their everlasting support, and for teaching me to choose my education out of interest.

To my brother Rasmus, thanks for our shared effort in learning about graphics during our Master's studies, and to him, and to Ulrik Lund Olsen, thanks for many good suggestions as to improvements of this thesis.

Thanks to my Ph.D. advisers Niels Jørgen Christensen and Peter Falster for their help in getting a Ph.D. scholarship for me, for their unfailing support throughout my studies, and for their mild guidance such that I was always able to pursue my own ideas.

A special thanks to Andreas Bærentzen for his open door policy, and for our numerous discussions on many aspects of the work presented in this thesis (I hope that I did not take too much of your time).

I would like to thank the advisers of my external stays. Henrik Wann Jensen for suggesting that we write a SIGGRAPH paper based on my work, and for the great help and hospitality given to me by him and his family during my stay in San Diego in January 2007. Geoff Wyvill for sharing his profound knowledge about noise and geometry, and for welcoming my three months visit at the University of Otago, New Zealand, in 2006.

Thanks also to Gert L. Møller for taking an interest in my work, and for co-authoring some of the first papers published during this project.

Other people who have had a positive influence on my work, and who I would

therefore like to thank, are Bent Dalgaard Larsen, Bjarke Jacobsen, Lars Schjøth, Henrik Aanæs, François Anton, and Anders Wang Kristensen. Everyone in the Lunch Club deserve a special thank you for making lunchtime more interesting and more fun.

I would also like to thank our librarian Finn Kuno Christensen who found paper copies of several old references that I was unable to locate myself.

For monetary support, I thank Otto Mønsteds Fond and AEG elektronfonden.

Parts of this thesis are based on a paper to be published and on a paper previously published. Although I was the primary author and investigator of these two papers, I would like to acknowledge the assistance and advice provided by the co-authors. The two following paragraphs provide references to the papers, and describe how they have been used in this thesis.

Chapter 2 is an extended version of the paper: Jeppe Revall Frisvad, Niels Jørgen Christensen, and Peter Falster. The Aristotelian rainbow: From philosophy to computer graphics. In *Proceedings of GRAPHITE 2007*, ACM, December 2007. To appear.

Sections 9.2 and 9.3 and Chapter 10 are, in part, based on the article: Jeppe Revall Frisvad, Niels Jørgen Christensen, and Henrik Wann Jensen. Computing the scattering properties of participating media using Lorenz-Mie theory. *ACM Transactions on Graphics*, Vol. 26, No. 3, Article 60, July 2007. Chapters 15, 16, and 17 are extended versions of the examples which appear in the same article.

Contents

Chapter 1

Introduction

Every new artist, and for that matter every new composer, is a problem child — a composite of virtues and defects that challenges the keenness of mind of the listener.

Aaron Copland, from *Music and Imagination*

Light is what you sense.
Matter is what you see.
Geometry is an abstraction over the shapes that you see.
Appearance is a combination of the three.

This thesis is about mathematical modelling of appearance for synthesising realistic images. It is also about analysis of the properties of materials using appearance models.

In the midst of three quite general research fields, namely physics, mathematical modelling, and computer science, we find this curious line of research which is computer graphics. The branch of graphics concerned with realistic image synthesis relies heavily on the first two fields, but could not exist without the third. So any person writing a thesis within this branch of graphics is faced with the problem of choosing which one of the three fields to put the emphasis on in the text. In this thesis mathematical modelling is in focus. You will encounter an abundance of physical theories throughout the thesis, but their purpose is to help finding mathematical models for realistic image synthesis. The models lead to algorithms which we can implement on a computer. Implementation details will not be in focus. Figure 1.1 is a schematic overview of realistic

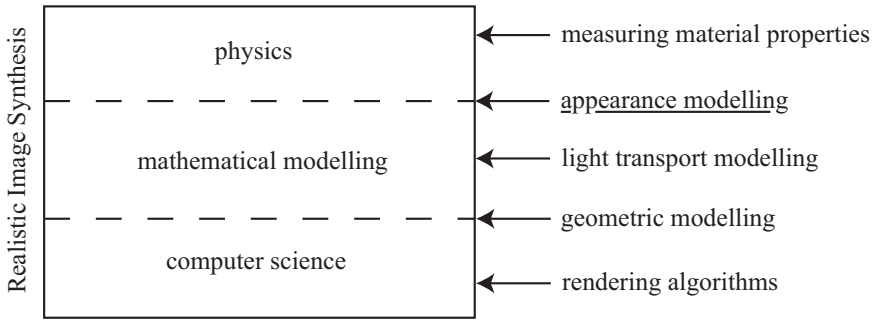


Figure 1.1: *A schematic overview of the branch of graphics known as realistic image synthesis. The list of subjects to the right is not complete. Many other subjects could be added. Appearance modelling has been underscored because it is the subject of this thesis.*

image synthesis. The subject of appearance modelling has been placed on the border between mathematical modelling and physics. This is not to say that appearance modelling does not involve computer science. It is to say where emphasis is usually placed in literature concerning the subject.

The process of synthesising an image from a model is known as *rendering*. An appearance model is then a model for realistic rendering. A large group of literature in realistic image synthesis is about doing realistic rendering efficiently. Appearance modelling is, on the other hand, about making the materials in a rendering look right. Therefore you will see only few comments with respect to rendering speed in the chapters to follow.

One way to do appearance modelling is to fit a model to measured material properties. The approach in this thesis is different. Instead of seeking the connection to the physical world through measurements, the approach taken here is to seek the connection between graphics and theoretical physics.

When a mathematical model is constructed using intuitive arguments about the phenomena that we observe in the world, it is said to be *phenomenological*. Such models typically work at a macroscopic level without any direct connection to the microscopic physical nature of the phenomena. Most models used in graphics are phenomenological. Indeed the fundamental rendering equations (both the one for surfaces and the one for volumes) are phenomenological. The exactitude of the *macroscopic material properties* which appear in these equations is all-important for capturing the right appearance of materials.

There are two ways to obtain macroscopic material properties (or three ways if

we include manual adjustment). Either we use measurements or we use computation based on an underlying theory. Since the connection to the microscopic physical models is not so obvious, the method of choice (in graphics) has been to measure the macroscopic properties. Unfortunately we need a lot of measurements if we measure macroscopic properties of materials directly. A connection to the microscopic level allows us to use a smaller number of measurements, and to combine them into a large number of different macroscopic properties for different materials. This is perhaps the most important reason why the connection between graphics and physics is important. Another reason is that the connection reveals the limitations of the phenomenological models. For these reasons, a large part of the thesis is devoted to the connection between macroscopic phenomenological theories of light and microscopic physical theories of light.

Measurements at a macroscopic level are still important. They should be used if we know exactly what type of material we want to render, and if we have a copy available for measurement. The connection to the microscopic level provides an entirely different type of appearance model. It provides a highly versatile model with many adjustable parameters. Thus models based on measurements and models based on theoretical development are complementary, not competitors.

The many adjustable parameters in a microscopic description of materials means that a number of choices must be made before we compute the macroscopic properties. We often have to choose whether a parameter should be included in the model or neglected, or whether an average value should be used. In a sense we have to limit the variability of nature. This high variability is unfortunate in practice, where we want the model to be simple and manageable. On the other hand, it is interesting in theory because we can learn about the visual variations which are due to many different details in the composition of a material.

Suppose we think of an appearance model as a multidimensional shape. Each dimension of the model is due to an adjustable parameter. If we find a way of representing the multidimensional shape, we can pick different slices of the shape to obtain practical models, and we can still retain the variability in the original shape for further studies. This thesis includes a proposal for a new way of representing multidimensional shapes. An important advantage of this geometrical representation of appearance models is that there is no distinction between input and output variables. They are all just another dimension of the shape. This makes it possible to use the multidimensional shapes not only for synthesis of realistic images, but also for analysis of the properties of a material based on its appearance.

Light, matter, and geometry are the essential elements in the construction and rendering of an appearance model. The interaction of light and matter gives

rise to the macroscopic optical properties for materials. Geometry is used to give the materials different (multidimensional) shapes. Different materials of different shapes comprise a scene. We use the optical properties and the geometry to compute how light propagates through a scene in a rendering. If the optical properties are correct, and if the geometry is appropriate, the result is a realistic image. This thesis is in four parts. In the first three parts, each of the three essential elements are investigated thoroughly. In the fourth part, three appearance models are presented as to exemplify the general results found in the theory concerning the essentials. In the remainder of this chapter, Section 1.1 provides pointers to the background of appearance modelling in graphics and Section 1.2 provides an overview of the contents of the four parts.

1.1 Background

It is difficult to say precisely where appearance modelling originates. Indeed modelling the appearance of materials has been one of the key motivations from the very beginning of computer graphics. Appearance modelling, as we think about it today, arises from realistic image synthesis. The work by Hall and Greenberg [1983] laid a solid foundation for realistic image synthesis. With this well-developed model for rendering of realistic images, it is natural to think about development of physically-based appearance models. One of the first papers to focus on physically-based appearance modelling is that of Haase and Meyer [1992] which introduces Kubelka-Munk theory in order to model the appearance of pigmented materials.

Many papers [Hanrahan and Krueger 1993; Dorsey and Hanrahan 1996; Dorsey et al. 1996; Dorsey et al. 1999; Jensen et al. 1999, etc.], which I would categorise as papers within the line of research called appearance modelling, followed the paper by Haase and Meyer [1992]. These papers, and subsequent papers on appearance modelling, are all based on physical modelling of the properties of materials. However, when the materials are rendered, the phenomenological radiative transfer theory is employed. In my opinion, this renders the foundation of appearance modelling incomplete. The aim of this thesis is, therefore, to provide the connection between the physical theories of light and matter and the phenomenological theories used in realistic image synthesis. Many appearance models are concerned with corrosion and other changes in appearance over time. This aspect of appearance modelling will not be considered. The focus of this thesis is the optical properties of materials.

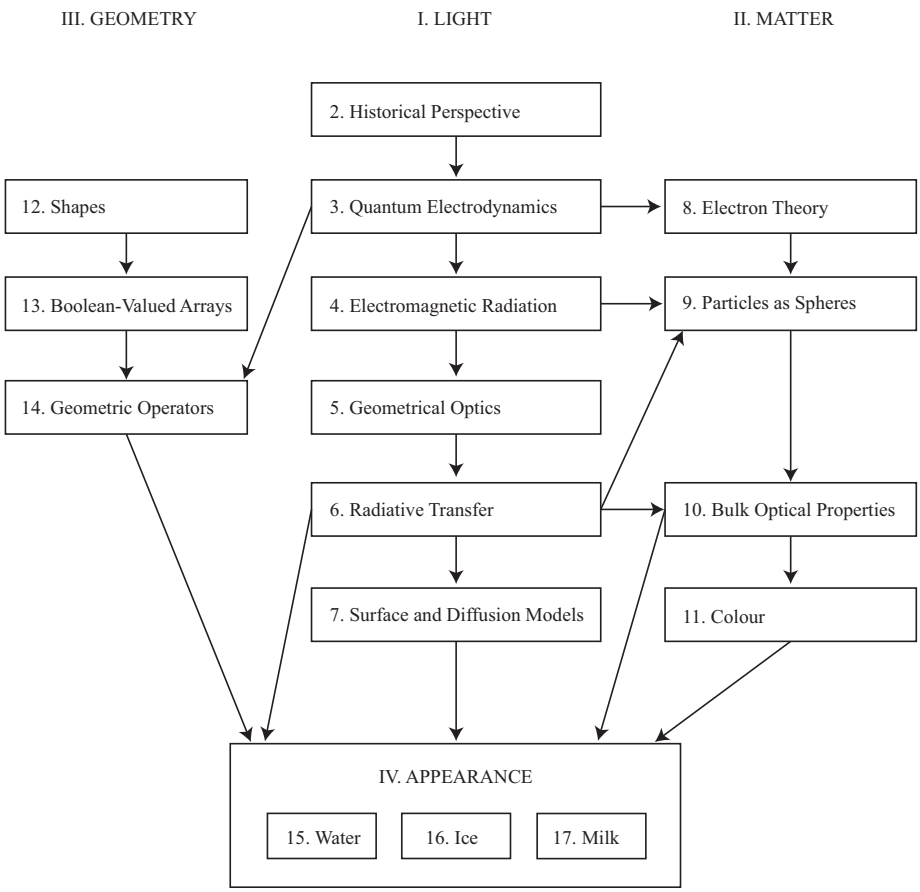


Figure 1.2: *A rough sketch of the connection between the theoretical contents of the different chapters. An arrow means that some theoretical contents of a chapter is used in the chapter the arrow is pointing at.*

1.2 Overview of the Contents

Figure 1.2 is a sketch of the connection between the theoretical contents of the different chapters. It should only be taken as a rough guide. The arrows in the figure are not strict rules. Another way to think of the figure is as follows: One may find it difficult to understand a chapter (or part of it) if the chapter(s) pointing to it have not been read. If the intension is not to read the thesis from a to z, Figure 1.2 may help identifying the chapters that one should read.

Part I is about light. The overall purpose of this part is to provide (a) the

connection between microscopic and macroscopic theories of light, (b) the connection between physical theories of light and the theories used in graphics, and (c) a general understanding of theories of light and their historical origins. Special emphasis is put on three aspects: (a) the origin of the macroscopic optical properties of materials in the microscopic physical theories, (b) the limitations of the theories of light that are most commonly used in graphics, and (c) the role of absorption and the complex index of refraction in optical theories of light propagation. We start from the very beginning with a historical perspective on theories of light and realistic image synthesis. The history of light is carried all the way up to quantum electrodynamics. Then we move in another direction. We start from quantum electrodynamics and make simplifications all the way down to the theories of light used in rendering algorithms for realistic image synthesis. Part I is in six chapters. The following six paragraphs describe the contents of each chapter.

Many theories of light have been proposed over the centuries. In graphics, we use many laws and concepts that have an early origin in the history of theories of light. To provide an overview, and to introduce the different laws and concepts in a verbal fashion, Chapter 2 pinpoints historical events in the development of theories of light with relevance for graphics. It is, to my knowledge, the first source-based study of its kind. The chapter has two key observations. It is observed that realistic image synthesis has developed in a way somewhat analogous to the development of the theories of light. Using the Aristotelian rainbow theory as an example, it is also observed that ancient theories of light still have their uses in graphics. As an introduction to the subject of quantum electrodynamics, the historical development is carried all the way up to this theory which describes all known aspects of light's behaviour.

Even if quantum electrodynamics is about light at the most microscopic level, it is interesting and important (in my opinion) to learn about the true nature of light. Chapter 3 is an introduction to quantum electrodynamics with emphasis on the connection to Maxwell's equations, and on the concepts that could lead to a simplistic quantum field simulator for graphics. The chapter also introduces some quantum mechanical concepts that will be used to describe the microscopic properties of materials in Part II. In addition, the use of operators for expansion and reduction of the degrees of freedom in a system of quantum particles is analogous to our treatment of multidimensional shapes in Part III. The connection to Maxwell's equations highlights the simplifications that are imposed on light when we go from a quantum theory to a wave theory.

Many chapters in this thesis are concerned with subjects that are based on Maxwell's electromagnetic field theory. Chapter 4 is a review of all the quantities and concepts from the electromagnetic field theory which are used subsequently in the thesis. Special emphasis is put on the introduction of optical properties

for materials and on the wave-theoretical justification of the laws and formulae that we use in graphics. The optical properties introduced in electromagnetic field theory connect the Maxwell equations which are used at a microscopic level to the Maxwell equations which are used at a more macroscopic level. Special attention is given to the complex index of refraction which sums up all the material properties introduced in electromagnetic field theory. The imaginary part of the complex index of refraction is often neglected in graphics. The imaginary part is, however, related to the absorption of a material, and most materials exhibit some absorption in the visible range of the electromagnetic spectrum. Therefore some effort has been spent throughout the thesis to discuss the imaginary part of the refractive index.

Graphics is mostly concerned with the visible part of the electromagnetic spectrum. The wavelength of visible light waves is very short. Therefore a ray theory of light is not such a bad approximation in graphics. The ray theory should, however, model the wave theory as accurately as possible. Geometrical optics is a ray theory of light which is exact in the limit where the wavelength is zero. Thus it is a good theory for graphics. A general theory for handling heterogeneous, absorbing materials in geometrical optics (homogeneous and non-absorbing materials are special cases) is developed in Chapter 5. A proposal is made for a new way of tracing rays that follow the energy propagation through absorbing materials more closely. The influence of absorption is most significant for objects which are small enough to let light reemerge on the surface only partly absorbed. Therefore light is traced through a small absorbing particle to illustrate the effects captured by the new way of tracing rays. Particles are essential in the link between the scattering of materials with one continuous phase and the macroscopic scattering properties used to describe turbid materials.

The macroscopic optical properties used in graphics are the complex index of refraction and the macroscopic scattering properties used in radiative transfer theory. Since the radiative transfer theory is phenomenological, it is not obvious how the scattering properties connect to the microscopic phenomenon of particle scattering. This connection is outlined in Chapter 6. It is an important connection because it allows us to compute the scattering properties of materials as described in Part II. The connection also reveals the limitations of the phenomenological models. At this point we have reached the macroscopic level where rendering is usually done. So a short account of volume rendering is also provided in this chapter.

Many rendering algorithms are concerned with surfaces rather than volumes. To complete the journey from quantum electrodynamics to rendering algorithms, the theory for volumes is connected to the theory for surfaces in Chapter 7. Diffusion-based rendering algorithms are particularly popular at the moment. To illuminate the limitations involved in diffusion-based rendering, a quite gen-

eral derivation (which I have seen nowhere else) of Fick's diffusion law is presented. The theory is coupled to the popular dipole approximation for subsurface scattering, and a suggestion for improvement is made. This chapter concludes Part I.

Part II is about matter. Just like there is a connection between the microscopic and the macroscopic theories of light, there is also a connection between the microscopic behaviour of matter and the macroscopic material properties used in a rendering. In this part the connection is outlined from the properties of atoms all the way up to the macroscopic optical properties used in rendering. Part II is in four chapters. The following four paragraphs describe the contents of each chapter.

The optical properties of atoms are a result of the quantum mechanical behaviour of electrons which are bound to nuclei. Slightly more macroscopically, we can think of atoms as damped harmonic oscillators with several different resonant frequencies. Chapter 8 is a review of the connection between the resonant frequencies of atoms and the complex index of refraction. With a limited set of properties for each atom, it is possible to compute the complex index of refraction approximately. The quantum mechanical justification for the blackbody emission spectrum is also briefly summarised.

The connection between complex indices of refraction and the scattering of an electromagnetic wave by a particle is a complicated subject. Especially if the particle is embedded in an absorbing host medium. This is very often the case, so it is important in a graphics context that we are able to handle an absorbing host. Previously only homogeneous waves have been considered for particles in an absorbing host. This is a little unfortunate because light waves are almost always inhomogeneous when they propagate through an absorbing medium. Scattering of inhomogeneous waves by a spherical particle in a *non*-absorbing host has recently been considered in the literature. In Chapter 9, the full computation for the scattering of inhomogeneous waves by a spherical particle in an absorbing host is presented. Even for homogeneous waves scattered by a particle in an absorbing host, evaluation of the theoretical solution has previously been numerically unstable. A new scheme for robust numerical evaluation of the theoretical solution is also provided. Finally, some considerations are made regarding non-spherical particles.

Particles often appear in a wide variety of sizes within a material. Chapter 10 is about finding the scattering of a cloud of particles, or, in other words, it is about finding the bulk optical properties of a material. The particles are assumed to be randomly distributed and not too densely packed. Within these limitations, it is described how to find the macroscopic optical properties of scattering materials. Different continuous distributions of particle sizes are considered.

A common approach in graphics is to use trichromatic colour values when specifying materials in a rendering. Chapter 11 is a compact review of the transition from spectral optical properties to representative trichromatic colour values. The physical theories described in the previous chapters of the part are only concerned with spectral optical properties. This chapter concludes Part II.

Part III is about geometry. The purpose is to develop highly versatile but also practical appearance models. The physical shape of a material is but an extra set of input and output to be part of an appearance model. There is no reason why we should not be able to treat both physical shape and all the adjustable parameters in an appearance model as the geometry of a multidimensional shape. The advantage of such an approach is that the multidimensional shape is a sort of database comprising all the properties of an object. Both optical properties, physical shape, and all the input parameters. Such a database enables us to draw new conclusions about a material. For example to find the relation between the optical properties of a glass of milk and the fat contents of the milk (regardless of the other input parameters). The multidimensional shape also makes it easier to animate an object by taking a slice in the shape. A slice could, for example, be the change in the appearance of water as it freezes. Part III is in three chapters. The following three paragraphs describe the contents of each chapter.

The idea of multidimensional shapes is introduced as a natural extension of the normal conceptions about shape and geometry. Chapter 12 is a brief overview of conventional representations of geometry in graphics.

To handle and represent multidimensional shapes, I have chosen to use array theory. It is a discrete mathematical theory which is convenient for handling problems of many dimensions. Since array theory is not so widely known, a short introduction is provided in Chapter 13. Although it is a discrete theory, it is conceptually straightforward to use it with shapes on continuous domains. However, to make it practical, we need a discrete representation. Discrete representations of continuous geometry are well-known in graphics. There are many different representations to choose from. Inspired by parametric surfaces, the concept of polynomial arrays is introduced in this chapter as a discrete representation of multidimensional shapes.

Expansion and reduction of dimensionality are the fundamental concepts in the handling of problems with many degrees of freedom. These are the concepts used, for example, in quantum electrodynamics, and they are what we use to draw conclusions from multidimensional shapes. Array-theoretic definitions of these operations are provided in Chapter 14. This chapter concludes Part III.

Part IV is about appearance. The purpose of this part is to exercise the theory

presented in the parts that precede it. Three examples are constructed: An appearance model for water, one for ice, and one for milk. The models for water and ice are not only for clean water and ice, but also (and more interestingly) for the particulate water and ice that we find in nature. There is a chapter for each of the three examples. The following three paragraphs describe the contents of each chapter.

The water example is the simplest example of the three. It is described in Chapter 15 which also provides an outline of the general procedure for computing optical properties of materials, and for constructing appearance models. Ample attention is given to the optical properties of pure water because pure water is an essential component in many natural materials. The appearance model, which is found in this chapter, enables us to compute the appearance of water as a function of temperature, salinity, and mineral and algal contents of the water.

The ice example described in Chapter 16 is very thorough (compared to the simpler water example), and it includes much information about the optical properties of freezing sea water. The main purpose of this chapter is to show that the presented procedure works well for a solid material which also contains non-spherical particles. The resulting appearance model is parameterised by temperature, salinity, and density of the ice. As for the water example, it is also possible to include the mineral and algal contents of the ice in the model. Some experimentation with mineral and algal contents is carried out to shed some light on the occurrence of green icebergs in nature.

Milk is interesting because its macroscopic optical properties have previously been measured in graphics. This gives us an opportunity to do analysis of the contents of the previously measured milk samples using an appearance model. In Chapter 17 an appearance model parameterised by fat and protein contents is constructed for milk. The prediction by this model of optical properties for different types of milk is used to estimate the fat content of measured optical properties. The reason for deviations in some of these estimates is analysed using the concept of multidimensional shapes.

1.3 A Note on Notation

A wide variety of theories are employed in this thesis. I have strived to follow the conventional symbols and notation of the different theories. This makes it easier to read the text for people who are acquainted with the theories. On the other hand it has the consequence that the same symbol sometimes is used to denote different quantities. Sometimes two different symbols are used to denote the

same quantity in two different contexts (to avoid ambiguities in one of them). As I see it, reuse of symbols has been inevitable. To avoid any major confusion, a symbol which is not explained immediately before or after it is used has the meaning that was last assigned to it in the text.

Some general conventions have been chosen. An arrow above a symbol (for example $\vec{\omega}$) means that it is a vector of unit length specifying a direction. Symbols in bold font denote a vector quantity which is not necessarily of unit length. A symbol with a hat (for example $\hat{H}$) denotes an operator. Array-theoretic operations are written in roman font in equations, but in bold font when written in a sentence. Array-theoretic operators are written in capitalised roman font both in equations and in sentences. In general, primes do not denote derivative. The only exception to this clause is the definition of the Lorenz-Mie coefficients in Chapter 9, where primes are used to avoid cluttering up the notation. Here it is, of course, stated explicitly that the primes denote derivative.

Some strange switches between notation may be experienced. Especially in Chapter 6, where notation switches from physics to graphics notation. The two symbols $\mathrm{d}\Omega$ and $\mathrm{d}\omega$ are both used to denote an element of solid angle.

Finally, a short comment on terminology: I use the two terms "index of refraction" and "refractive index" interchangeably. In the context of array theory I distinguish between the terms "operation" and "operator" as follows. An operation is a function which takes data (for example numbers or arrays) as arguments and returns data. An operator is a function which takes an operation as argument and returns an operation. The term "operator" is sometimes used differently in other mathematical disciplines. Therefore the use of the term "operator" will not be consistent throughout the thesis, only in the context of array theory.

Part I

LIGHT

Chapter 2

Historical Perspective

History unravels gently, like an old sweater. It has been patched and darned many times, reknitted to suit different people, shoved in a box under the sink of censorship to be cut up for the dusters of propaganda, yet it always - eventually - manages to spring back into its old familiar shape. History has a habit of changing the people who think they are changing it. *History always has a few tricks up its frayed sleeve. It's been around a long time.*

Terry Pratchett, from *Mort*

On immediate inspection the ancient theories of light would seem to have very little to do with the very modern phenomenon of synthesising life-like images on a computer screen. Nevertheless, we shall explore in this chapter how the development of various theories of light has many things in common with the development of algorithms for producing photo-realistic computer imagery.

Realistic image synthesis is a research field which appeared at a very late point in history. At least this is true if we exclude paintings and only consider images rendered on a computer. At the time where this branch of computer graphics emerges, the quantum theory of light is able to explain every known detail of light's behaviour. And the behaviour of light is exactly what we need to simulate if we are to compute the appearance of scenery. From the outside it may then seem like a paradox that graphics research has never used the most exact theory of light. Here is the reason why: the history of theories of light embraces a long period of time in which physicists have strived to understand nature in increasingly fine detail. The finer the detail, the more complicated the

Figure 2.1: *A rainbow rendered in real-time using Aristotle's theory for rainbow formation.*

complete picture. However in order to do computer graphics, we have to model the complete picture. There is no way around it, and it will probably result in a very crude model, but for every object in an image and every source of light we need a mathematical model to start from. Creating a realistic image from these artificial models quickly becomes an immensely complicated thing to do. If we want to see the result before the computer melts down, we have to start with a simple theory of light. Thus as computers grew more powerful, graphics research incorporated more and more detail and developed in a way somewhat similar to the development of theories of light.

To follow the development of theories of light, let us begin by looking at the first texts that have survived, in a more or less corrupted version, from ancient times (Sec. 2.1). Without doubt there have been theories of light before these, but many manuscripts have been lost [Smith 1999]. There are actually many parallels between this first groping towards an understanding of vision and the most common rendering algorithms for computer graphics. From antique theories we move on to more recent wave and radiation theories (Sec. 2.2). These, especially radiative transfer theories, are being used more and more often in graphics. Then I give a short account of the developments in realistic image synthesis (Sec. 2.3) and so as to exemplify how we can exploit the insight that realistic rendering is related to theories of light, I demonstrate that Aristotle's theory of rainbows provides an easy way to render rainbows in real-time (Sec. 2.4). The result of such a rendering is shown in Figure 2.1. To give a feeling of what may await in future graphics research, the historical development of quantum theories is also covered (Sec. 2.5). This thesis is inspired by the theories of light that have not yet been considered much in graphics. Let us try to push the development of algorithms for realistic rendering by studying the more detailed theories of light from a graphics perspective.

2.1 Ray Theories

The stories about the early Greek philosophers compiled by Diogenes Laërtius [∼200 A.D., 1901], provide an opportunity to get an understanding of the philosophy leading to the first theories of light. Reading Laërtius' account of the theories of Pythagoras (c. 575 – c. 495 B.C.), we understand that the light from the sun was thought of as a source of heat and life rather than a direct cause of human vision. Laërtius writes that one of Pythagoras' theories was [Laërtius ∼200 A.D., 1901, p. 349]:

> that the sun, and the moon, and the stars, were all Gods; for in them the warm principle predominates which is the cause of life. [...] Moreover, that a ray from the sun penetrated both the cold aether and the dense aether, and they call the air the cold aether, and the sea and moisture they call the dense aether. And this ray descends into the depths, and in this way vivifies everything.

Laertius explains Pythagoras' theory of the senses on this basis. Since man is alive, he contains the warmth received through rays of light from the sun. By emanating vapour of excessive warmth from the eyes, he is allowed to see through air, and through water. Laërtius tells us that Pythagoras "calls the eyes the gates of the sun".

This very early theory describes an indirect relation between light and sight (and heat). If there is no light, we do not receive heat and consequently have no excess warmth by which we can gather impressions using our eyes. Since everything which emits light also emits heat, it is easier to understand why it took several centuries before it was finally concluded that vision is not caused by rays from the eyes. This does not mean that the ancient Greek philosophers did not discuss the possibility of the eye playing only a passive role as a receptor of visual impressions. Very early on, such a theory was put forth by Democritus (c. 460 – c. 375 B.C.). According to Theophrastus (c. 371 – c. 287 B.C.), Democritus explains vision by a reflection or image in the eye as follows [Theophrastus ∼300 B.C., 1917, §§50–51]:

> the air between the eye and the object of sight is compressed by the object and the visual organ, and thus becomes imprinted (*typousthai*); since there is always an effluence of some kind arising from everything. Thereupon, this imprinted air, because it is solid and of a hue contrasting [with the pupil], is reflected in the eyes, which are moist. [...] Democritus himself, in illustrating the character of the "impression", says that "it is as if one were to take a mould in wax".

What Democritus describes seems most of all akin to a mechanical process. His description does not involve light. It is more like "a sort of stamping-process, the result of which can be seen in the images reflected at the cornea's surface" [Smith 1999, p. 25].

This explanation of vision was not Democritus' own personal opinion, but rather Democritus' way of describing the supposition of a number of presocratic thinkers referred to as the natural philosophers. To counteract it, Plato (c. 427 – c. 347 B.C.) and Aristotle (384 – 322 B.C.) had to go through some trouble to explain why light plays a part in the workings of vision (see for example Plato's *The Republic*, 507c–508b). Plato was influenced by the Pythagoreans and Aristotle was a student of Plato. So in a way they developed the thoughts of the Pythagoreans to a more advanced state. The argument of Plato and Aristotle as to why light must have a role to play in vision, is really quite simple. Essentially the argument is that we can look at a colourful object and even if we do not change our way of looking, the object can still loose its colour. Conclusively there must be a third thing, outside object and eye, influencing our ability to see. This third thing is, of course, light.

In his later work, Plato presents a theory of vision which is a pleasant compromise between the previous theories. In *Timaeus* [∼360 B.C., 1989, 67c] he writes that "colours [...] are a flame which emanates from every sort of body, and has particles corresponding to the sense of sight". This is quite analogous to the account of Democritus, but he also gives the following account of how vision works [Plato ∼360 B.C., 1989, 45b–45d]:

> When the light of day surrounds the stream of vision, then like falls upon like, and they coalesce, and one body is formed by natural affinity in the line of vision, wherever the light that falls from within meets with an external object. [...] But when night comes on and the external and kindred fire departs, then the stream of vision is cut off; for going forth to an unlike element it is changed and extinguished, being no longer of one nature with the surrounding atmosphere which is now deprived of fire: and so the eye no longer sees, and we feel disposed to sleep.

The interesting development is that light is more directly involved in the process in Plato's account of vision. It is also interesting to note that Plato refers to a stream of vision as "the light from within". The meaning of light is changing. It is no longer only thought of as the life-giving fire emanated from the sun. Aristotle makes this change of conception more clear by saying that light is not an emanation from the sun or the eye, but rather an instantaneous thing which exists when the potentially transparent (e.g. air and water) is actually transparent (or "is excited to actuality" as he puts it) [Aristotle ∼350 B.C., 1941, II:7].

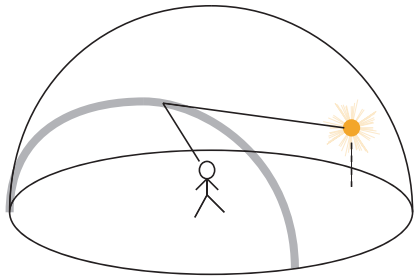


Figure 2.2: *An illustration of Aristotle's rainbow theory. Clouds on a hemisphere resting on the circle of the horizon reflect sunlight to the observer where the angle is equal (to some constant angle).*

What we have discussed so far are the different theories of the antiquity which try to explain the physics behind visual perception. It is evident that at this point the concept of light is in a far too fuzzy state to enable mathematical treatment of the subject. The theory of vision is, however, an entirely different matter. It is easy to follow the line of sight and realise that we can describe it in a mathematical way. This is the subject of *optics* which was already well developed at the time of Aristotle [Smith 1999]. Book III of Aristotle's *Meteorology* [∼350 B.C., 1984] demonstrates quite advanced thoughts. He explains the appearance of halos and rainbows by considering reflection of the line of sight in mist around the sun (for halos) and clouds just before they condense into rain (for rainbows). After arguing that these phenomena are the result of reflection, he uses the idea that the angle between the line of sight and the direction from the cloud to the "luminous body" should be equal wherever the rainbow or halo is seen. Using this principle and placing clouds on "a hemisphere resting on the circle of the horizon", he is able to explain the appearance of halos and rainbows in a mathematical way. Figure 2.2 illustrates the rainbow theory.

Unfortunately no manuscripts have survived from the initial phase of optical studies. Euclid's *Optics* [∼300 B.C., 1945] and *Catoptrics* [∼300 B.C., 1895] are the oldest surviving works dedicated entirely to optics. The principles of *perspective* are established and the perceptual distortions resulting from our conical vision (as well as a few propositions on binocular vision) are considered in the *Optics*. The *law of reflection* is the first proposition of the *Catoptrics* [Euclid ∼300 B.C., 1895, p. 287]:

> A speculis vel planis vel convexis vel concavis radii sub angulis aequalibus refringuntur.
> (Translation: *Rays are reflected at equal angles by plane, convex, and concave mirrors.*)

(Modern version [Saleh and Teich 2007, p. 5]: *The reflected ray lies in the plane of incidence; the angle of reflection equals the angle of incidence.*)

Using this proposition a number of effects resulting from reflections in concave and convex mirrors were derived by Euclid.

With optics at such an advanced state only relatively shortly after Aristotle's *Meteorology*, we might wonder whether Aristotle knew about the law of reflection or not[1]. He does use a principle of equal angles as the key to describe halos and rainbows by reflection off mist and clouds (respectively), but on first inspection it seems different from the Euclidean proposition. I do not believe it is a different principle. Suppose Aristotle's derivations are based on the fact that the orientation of the cloud surface, where the visual ray impinges, is unknown. This is a perfectly reasonable assumption because his writings very elegantly avoid having to specify this orientation. What we have to assume then, to make the proofs work with the correct law of reflection, is that the cloud (or mist) surfaces have the same orientation towards the sun across the entire hemisphere. If this is kept in mind when reading Book III of *Meteorology*, the explanations make a lot more sense (in my opinion). That the cloud surface is able to exhibit this unusual behaviour is explicable by its particulate nature, the same nature which Aristotle uses to explain why we do not see a perfect reflection of the sun in the cloud.

Although optics started out being a theory of vision rather than a theory of light, developments took an interesting turn when the mathematicians took an interest in burning mirrors. The earliest known work on this subject is Diocles' treatise *On Burning Mirrors* [∼190 B.C., 1975]. It treats the focusing properties of parabolic sections. This means that it had been realised that rays of light from the sun follow the same general rules as rays of sight from the eyes. The theory of light is getting less fuzzy. By applying the rules of optics to rays of light, it can be established that light moves in straight lines and that the law of reflection is also valid for rays of light.

The supposition that visual flux issues forth from the eyes persists. But in light of Diocles' work, we are allowed to believe that what is true for rays of sight is equally true for rays of light. A few centuries later, a new insight into the behaviour of rays is obtained by Hero (or Heron) of Alexandria. In his *Catoptrics* [∼50 A.D., 1900] Hero uses an arrow as an example and says that "because of the impelling force the object in motion strives to move over the shortest possible distance, since it does not have the time for slower motion, that is, for motion over a longer trajectory. And so, because of its speed, the object

[1] Some authors, e.g. Smith [1999], say that Aristotle's writings violate the law of reflection, but this is not necessarily so.

tends to move over the shortest path" [Smith 1999, §2.6]. Then he observes that "the rays emitted by us travel at an immeasurable velocity" as "it takes no perceptible time for [them] to reach the heavens". The implicit conclusion is that as rays travel at an almost infinite speed, they do not only strive to take the shortest path, they have to take it. This principle that rays take the path of minimum distance is now known as *Hero's principle*. Hero himself uses it in his *Catoptrics* to demonstrate the law of reflection.

With respect to refraction, Hero's *Catoptrics* [∼50 A.D., 1900] only attempts an explanation of why light is partially reflected and partially transmitted into water and glass. Another century had to pass before proper treatment of refraction was to be presented by Ptolemy (c. 100 – c. 178 A.D.) in his *Optics* [∼160 A.D., 1996]. From a physical point of view, the work of Ptolemy is immensely important because he makes extensive use of carefully contrived experiments to support his arguments. This signals the beginning of a new era where pure philosophical reasoning is not necessarily the main authority. Ptolemy's measurements of the angle of refraction are surprisingly exact. He found, in Book V of the *Optics*, the angle that rays make when moving "from rarer and more tenuous to denser media" (i.e. from air to water to glass) and the other way around, and he was able to describe this behaviour qualitatively, however he did not succeed in formulating the mathematical law of refraction. Ptolemy's *Optics* contains many fine results and we can certainly think of it as the culmination of ancient mathematical optics.

Ptolemy also writes qualitatively about the *shading* of objects as depending on the angle of the incident rays. In Book II of his *Optics*, he explains concepts which are surprisingly similar to diffuse and glossy reflections of light. He writes [Ptolemaeus ∼160 A.D., 1996, §§18–19 (pp. 76–77)]:

> everything that falls orthogonally strikes its subjects more intensely than whatever falls obliquely. Also, what is polished is seen more clearly than what is rough, because there is disorder in a rough object resulting from the fact that its parts are not arranged in a regular way. But the parts of a polished object have a certain regularity, and [so] brightness is inherent to it.

After Ptolemy the development in optics and theories of light is almost stagnant for several centuries. The only thing to mention is a description of *colour bleeding* by Galen (c. 130 – c. 200 A.D.). Colour bleeding is the phenomenon where light is tinted by the colour of nearby surfaces due to diffuse reflections. In the words of Galen [∼180 A.D., 1984, §7.7]: "when a person reclines under a tree [...], you can see the color of the tree enveloping him. And often when bright air touches the color of a wall, it receives the color and transmits it to another

body, especially when the wall is blue or yellow or some other bright hue". By "bright air" Galen probably means air excited by light.

Perhaps the downfall of the Roman Empire was the reason for the period of stagnation after Ptolemy. The prosperity of the Abbasid Caliphate which followed moved the scientific lead to the Arab world. In the second half of the 9th century, the Arabs start contributing to optics and theories of light. Ya'qūb al-Kindī (c. 801 – c. 873) addresses a subject which also troubled Ptolemy somewhat. Ptolemy had problems with the Euclidean idea that rays are distributed discretely in the visual cone. He opted that "the nature of the visual radiation is perforce continuous rather than discrete" [Ptolemaeus ∼160 A.D., 1996, II, §50]. Still he treated rays as "virtually discrete" [Smith 1999] such that he could follow their rectilinear trajectories in reflection and refraction. What Ptolemy was groping for, was the concept of solid angles and their differentials, but the mathematics available to him were not sufficiently sophisticated. The idea that radiation is spread in a continuum over solid angles is, however, important and it is further developed by al-Kindī. In his optical treatise [Al-Kindī ∼870, 1997] called *De aspectibus* in Latin, al-Kindī analyses the spread of radiation from a point source and writes: "what lies closer to [the] point is more intensely illuminated than what lies farther from it" [Smith 1999, p. 162].

After al-Kindī this development gathered momentum and Ibn Sahl (c. 940 – c. 1000) composed an impressive work *On the Burning Instruments* [Ibn Sahl ∼984, 1993]. In this he finds the *law of refraction*. If θ_1 and θ_2 are the angles formed by the normal of a plane surface and (1) the refracted light ray in a crystal and (2) the ray in the air, then [Rashed 1990, p. 478]:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{1}{n} \ ,$$

where n is the reciprocal of what we today would refer to as the relative index of refraction. With this law Ibn Sahl is able to couple the theory of burning mirrors as described by Diocles [∼190 B.C., 1975] with the theory of refraction as advanced by Ptolemy [∼160 A.D., 1996, V]. This led him to the first specifications of lenses.

Had the concept of lenses been in place, it might have been easier to comprehend that the eye really works as a passive sensor of light. Not long after Ibn Sahl's work on burning lenses the renowned Arab scientist Ibn al-Haytham (965–1039), known to Europeans as Alhacen, incorporated a peculiar type of lens in his model of the eye and dedicated the entire first volume of his *Kitāb al-Manāẓir* ("Book of Optics") [Ibn al-Haytham ∼1016, 2001] to the discouragement of the theory that vision issues forth from the eye (especially confer the conclusive line of arguments in §§6.45–6.60 of the reference). Even though Ibn al-Haytham opposes the theory of visual rays, he also explicitly makes it clear that all the

mathematical results involving rays of sight are still true, but in reality the rays consist of light travelling in the opposite direction. In the spirit of Ptolemy's *Optics*, Ibn al-Haytham's *Kitāb al-Manāẓir* comprises seven books covering all aspects of optics known at that time. He also carries Ptolemy's extensive use of experimentation further and uses it for verification of his theories throughout his treatise. Clearly al-Haytham's work is monumental in optics and upon its translation into Latin (c. 1200), it spawned renewed interest in the field.

Despite the efforts of al-Haytham, the new western scientific works on optics did not immediately discard the Greek tradition involving rays of sight. Perhaps the reason was the peculiar lens in al-Haytham's eye model, which he describes as only being sensitive to rays of light striking the surface of the lens orthogonally [Ibn al-Haytham ∼1016, 2001, I:7]. At the beginning of the 17th century, Johannes Kepler [1604, 2000] finally brought an end to the theories involving rays of sight. This was done by demonstrating that the lens of the eye is a perfectly ordinary lens merely serving the purpose of drawing an upside down image of what we are looking at point-by-point on the retinal screen behind the eye (the fact that the resulting image is upsidedown led al-Haytham to form his lens with special sensitivity). With Kepler's work the scene is set for further investigation into the nature of light.

Unfortunately the work of Ibn Sahl had not been translated into Latin, so the Europeans had to reinvent the sine-law of refraction. Kepler [1611] found an approximation of the law and discovered the existence of *total internal reflection*, which is the phenomenon that light cannot refract out of a dense transparent medium (e.g. glass) at a grazing angle, instead it will only reflect internally. According to Kwan et al. [2002], Thomas Harriot had already discovered the sine-law in 1602 and, likewise, Willebrord Snel van Royen (Latinised as Snellius) reinvented the law of refraction in 1621, but neither of the two published their results. The law was first published by René Descartes in his *Discourse on Method* containing a scientific treatise on optics [Descartes 1637, 2001]. Nevertheless the law of refraction is today called *Snell's law*. Descartes [1637, 2001, pp. 65–83] explains refraction by thinking of light as particles on which different friction-like forces act. The forces depend on the type of media which the particles are moving from and to. For Descartes' arguments to fit the experimental behaviour of light, he must draw the rather peculiar conclusion that light is received more easily by water (and even more easily by glass) than by air. To Pierre de Fermat this explanation was not convincing. Rather he felt that there should be a minimum principle from which the law of refraction can be derived [Fermat 1891–1912, pp. 354–359, letter from Fermat to De la Chambre, 1657], just like Hero used his principle of shortest path to derive the law of reflection. After putting his mind to it, Fermat [1891–1912, pp. 457–463, letter from Fermat to De la Chambre, 1662] found that he was able to derive the law of refraction from precisely such a principle. Conclusively he writes:

> my principle, and there is nothing that is as probable and as apparent as this proposition, [is] that nature always acts by the easiest means, or, in other words, by the shortest paths when they do not take longer time, or, in any case, by the shortest time

In short, *Fermat's principle* is that light follows the path of least time. This is a very powerful principle by which many things can be predicted, among them the laws of reflection and refraction. Today we know that light takes the path along which the time of travel is an extremum compared to neighbouring paths, and the wording of Fermat's principle has been adjusted accordingly in modern books on optics. The extremum is, however, usually a minimum and consequently Fermat's original formulation is true in most cases. In a homogenous medium (where the index of refraction is everywhere the same) the speed of light is the same everywhere. The path of least time is then also the shortest path. Thus Hero's principle is a special case of Fermat's principle.

This concludes our discussion of ray theories of light. Many of the ideas and principles from the two millennia of history that we have now discussed are indispensable in computer graphics today. They are the backbone of most photorealistic rendering algorithms. In particular we use: the fact that rays of light move in straight lines in homogeneous media; the laws of reflection and refraction; total internal reflection; the concepts of shading and colour bleeding; the concept that the energy carried by light is spread over solid angles; and Fermat's principle by which we are able to find the path of light in heterogeneous media (where the speed of light may change throughout the medium).

2.2 Wave and Radiative Transfer Theories

The work of Aristotle was widely read and quite influential in the seventeenth century [Shapiro 1973]. As mentioned previously, Aristotle thought of light as an excitation of potentially transparent media rendering them actually transparent. This idea caused many seventeenth century scholars to seek a theory in which light is a state propagating through a continuum (rather than particles following straight lines).

Inspired by Aristotle, Kepler [1604, 2000, Chapter 1] promotes the view that rays of light are merely a geometrical representation of what, physically, is a luminous spherical surface propagating from the centre of a light source. Following the same tradition, but carrying the concept further, Hobbes [1644, Prop. 4] writes "a ray is, in fact, a path along which a motion is projected from the luminous body, it can only be the motion of a body; it follows that a ray is the

place of a body, and consequently has three dimensions". Elaborating on the concept that rays of light are three-dimensional (parallelograms), Hobbes is able to derive the law of refraction without having to make the same counterintuitive assumption as Descartes (which was that light is easier received in glass than in air). Hobbes [1644] refers to the front of his rays as "propagated lines of light" and states that the width of the ray should be taken to be "smaller than any given magnitude" [Shapiro 1973]. This shows how remarkably similar Hobbes' concept of solid rays is to infinitesimal portions of an expanding wave. Hobbes also realises in a later version of his *Tractatus Opticus* that his ray concept is entirely different from traditional rays and then he introduces the new term *radiation* [Shapiro 1973, p. 151].

The works of Hobbes had considerable influence on subsequent theories of light. In an attempt to explain the *interference* colours of thin films, Robert Hooke [1665] proposes a peculiar mixture of Descartes' and Hobbes' theories of light. He uses Descartes' way of explaining refraction (and assumes that light moves faster in water than air), but he uses Hobbes' concept of solid rays, only he calls them light pulses. Hooke qualitatively arrives at the right conclusion about interference, namely that the colours of thin films are caused by reflection beneath the transparent film layer resulting in a delayed (weaker) pulse being "confused" with the pulse reflected at the surface. Hooke is, in other words, able to explain interference phenomena by treating the "propagated line of light" as the surface of constant phase. This clearly speaks in favour of a wave theory of light.

About the same time as Hooke investigates interference colours, Francesco Maria Grimaldi [1665, Book I, Prop. 1] observes that the path of light not only differs from a straight line when it is reflected or refracted, but also "when parts of light, separated by a manifold dissection, do in the same medium proceed in different directions". In other words, if you shine light at a very finely sliced object (a manifold dissection), you will observe light in the geometrical shadow. He calls this phenomenon *diffraction* and finds that it is best explained if light is thought of as a very fluid and very subtle substance.

Only a few years later Isaac Newton [1671] finds the correct explanation for the spectrum of colours seen when light is refracted through a prism. This phenomenon is called *dispersion* and it is due to the fact that, in the words of Newton [1671, p. 3079], "Light itself is a *Heterogeneous mixture of differently refrangible Rays*". In other words, Newton observes that each ray is disposed to exhibit only one particular colour and when rays of all the primary colours are mixed in a "confused aggregate of rays" light attains "whiteness" [Newton 1671, p. 3083]. Newton uses a cunning experiment to illustrate his theory. If he lets sunlight pass through a single prism, he sees a spectrum on the wall. But using a second prism after the first one, he sees light which is no different from the light coming directly from the sun.

During the same period of time, a strange phenomenon, which we today call *birefringence*, was discovered by Rasmus Bartholin [1670]. In his experiments with the crystal called Iceland spar, he observes that not only the ordinary image predicted by Snell's law, but also an "extraordinary" image is seen through the crystal. Bartholin regards this remarkable phenomenon to be a property of the crystal alone. It is, however, discovered a few years later that the experiment says quite a lot about the nature of light as well (see Huygens' discovery below).

Yet another property of light was ascertained in this period. While Descartes (in the Aristotelian tradition) was of the opinion that light is an instantaneous thing. Others such as Hobbes and Grimaldi (like Hero of Alexandria), were of the opinion that light travels at a finite, but imperceptible, velocity. By actually giving an empirically based estimate of *the speed of light*, Ole Rømer [1676] finally discounted the hypothesis that light is an instantaneous thing.

Many of the properties that have been discovered at this point (interference, diffraction, finite speed) lead towards a wave theory of light. In 1678 Christiaan Huygens completes his *Traité de la lumière* [1690] in which he presents a wave theory of light based on the theory of sound waves as it was known at the time. Huygens assumes that every particle of a luminous body emits a spherical wave. Moreover he enunciates the principle that every element of the wave fronts also gives rise to a spherical wave, and the envelope of all these secondary waves determines the subsequent positions of the wave front. This principle is today named after him [Born and Wolf 1999] and with it he is able to explain not only the laws of reflection and refraction, but also the extraordinary refraction in Iceland spar. However, letting light pass through a sequence of two Iceland spar, Huygens discovers that the waves of light change. They "acquire a certain form or disposition" [Huygens 1690, p. 94] because when the second crystal is in a certain position the two wave fronts emerging from the first crystal are not split again. In this way Huygens discovered *polarisation* of light, but he was not able to explain it theoretically.

With all these newly found properties of light and a wave theory ready for action, things take an unexpected turn. Enthused by his explanation for dispersion, Newton publishes his *Opticks* [1704] where he advocates strongly in favour of a ray theory of light. Two theories are then available: Newton's theory of "differently refrangible rays" and the wave theory of Huygens which explains birefringence. But being strongly in favour of a ray theory, Newton attempts, in a set of queries added in the first Latin version of the *Opticks* (1706), to give a ray-based explanation of birefringence and polarisation. His explanations are incorrect, but fact is that Newton's work became highly influential in the eighteenth century while Huygens' treatise was almost forgotten.

Concerning another aspect of light, namely how the intensity of light changes

under different circumstances as observed by Ptolemy (shading) and al-Kīndi (spread of radiation), there is no major development for several centuries. What is missing, in order to develop the subject quantitatively, are the means to measure the intensity of light. Such means are discovered by Pierre Bouguer in 1725 [Middleton 1964]. He invents a *photometer* by letting light from the source that he wants to measure the intensity of, fall on a screen. He compares this light to light falling on the same screen from a number of candles. By changing the distance of the candles to the screen, he is able to adjust the intensity due to the candlelight until it fits the other light. He uses the fact that the intensity of light is proportional to the inverse of the square of the distance from the source, and thus he is able to measure the light intensity in terms of candles. Bouguer's description of the technique is available in his *Essai d'optique* [1729]. I am uncertain whether Bouguer was the first to describe the fact that light falls off with the square of the distance to the source (the inverse square law for radiation), but he was the first to describe the formulae for finding the illumination I at a distance r from a source of intensity I_0 in a partially transparent medium [Bouguer 1729]. The modern form of the law is [Middleton 1964]:

$$I = I_0 r^{-2} e^{-\sigma_t r} \quad ,$$

where σ_t is the *extinction coefficient* of the semitransparent medium. If we consider a collimated beam of light the inverse square of the distance, of course, does not appear in the formula. The law stating the exponential falloff in intensity, $I = I_0 e^{-\sigma_t r}$, for a collimated beam, is often referred to as Beer-Lambert's law. The correct name is *Bouguer-Lambert's law.*

The contribution of Johann Heinrich Lambert [1760] to Bouguer-Lambert's law is that he gives it a mathematical formulation using logarithms. His *Photometria* [Lambert 1760] is an impressive work in this new field of research founded with Bouguer's photometer. Lambert [1760] also finds the *cosine law* which says that light reflected by a perfectly diffuse surface (also called a *Lambertian surface*) decreases in intensity with the cosine between the surface normal and the direction towards the incident illumination. This is the quantitative description of Ptolemy's observations about the shading of rough surfaces.

In the middle of the eighteenth century, but without reference to Huygens, Leonhard Euler [1746] gives a wave-based description of dispersion. This is accomplished by realizing that the colour of a light pulse is determined by its frequency. The next sign of weakness in the Newtonian ray theory appears in 1788 when René-Just Haüy investigates the birefringence of Iceland spar and finds a definite disagreement with Newton's results, but better agrement with the results of Huygens [Shapiro 1973]. When Thomas Young [1802] qualitatively explains the colours of thin films using the principle of interference between waves, the wave theory gets the upper hand. Especially as Huygens' explanation

of double refraction is confirmed by both William Hyde Wollaston [1802] and Étienne Louis Malus [1810]. Malus also discovers a previously unknown property of light which is that reflection causes polarisation.

With two very different theories striving for supremacy, the supporters of the Newtonian ray theory "proposed the subject of diffraction for the prize question set by the Paris Academy for 1818" [Born and Wolf 1999, p. xxvii]. To their dissatisfaction, the prize went to Augustin Jean Fresnel [1816] who, by an impressive synthesis of Huygens' envelope construction and Young's principle of interference, managed to overcome some theoretical difficulties in the previous wave theories and was also able to explain the diffraction phenomenon.

At the same time Fresnel was working on polarisation in cooperation with Dominique François Arago. They found that waves polarised at right angles to each other never interfere [Levitt 2000]. With this information Fresnel realised that the waves must be transverse rather than longitudinal and in 1821–1822 he presents three *Mémoires* [Fresnel 1827] in which he uses transverse waves to explain the birefringence and polarisation observed in crystals by Bartholin and Huygens. Shortly after these very strong arguments in favour of the wave theory of light (in 1823) Fresnel gives the ray theories the final blow: He presents formulae finding the intensities of the reflected and refracted waves and he even includes polarisation in these formulae [Fresnel 1832]. In this way he is able to explain Malus' observation that reflection causes polarisation. The *Fresnel equations*, as they are called today, are still used extensively.

At the beginning of the nineteenth century John Leslie [1804] firmly establishes that "heat and light are commonly associated". Thus when Julius Robert Mayer [1842] finds a relation between heat and mechanical energy and when James Prescott Joule [1843] subsequently discovers a similar relationship between heat and electromagnetic energy, there is "only" one step missing in the connection between heat, electromagnetism, and waves of light. This step is taken by James Clerk Maxwell [1873] who puts forth his famous theory of the electromagnetic field and gives substantial theoretical evidence to the fact that light waves are electromagnetic waves. His theory relies on the assumption that the speed of an electromagnetic wave, within experimental error, should be the same as the speed of light. Heinrich Hertz [1888] later verifies this conjecture by direct experiment.

Another theory was initiated at the beginning of the century when Young [1802] suggested that the eye most probably has three types of "fibres", each only sensitive to one of three different "principal colours". While first abandoned, this theory is revived by Hermann von Helmholtz [1867] who records three curves over wavelengths which represent the light sensitivity of each cone receptor in the eye. Each cone represents one of the three principal colours: Red, green,

and blue. The colours that we see are according to this theory (which is still generally thought to be true) a mix of the three principal colours weighted according to the wavelengths in the incident illumination. Today this theory of trichromatic colour vision is sometimes referred to as *Young-Helmholtz theory*.

Some years before Maxwell's electromagnetic field theory, Gustav Robert Kirchhoff makes an important observation when he investigates the connection between absorption and emission of a radiating body. Kirchhoff [1860, p. 277] introduces a hypothetical "completely black body", or for short a *blackbody*, which is a perfect absorber at all wavelengths. From the laws of thermodynamics Kirchhoff knows that if a body is kept at a constant temperature T, it will radiate as much energy as it absorbs. Using this fact, he proves that a blackbody will exhibit an emission spectrum which is determined by a universal function $J(\lambda, T)$. In a sense this function is the emission limit for all bodies at temperature T. If the body is not a blackbody, some wavelengths will be less intense than dictated by J, but the wavelengths (λ) at which the body will be able to emit are determined by the blackbody emission spectrum J. Kirchhoff [1860, p. 292] notes that it is of the utmost importance to find the function J, but that he finds big difficulties in determining it experimentally.

The first step towards finding the blackbody emission spectrum is taken by Jožef Stefan [1879] when he empirically finds that the heat radiated from a body is proportional to the fourth power of its temperature T. A few years later Ludwig Boltzmann [1884] gives the result a theoretical justification. The formula for a blackbody is today called the *Stefan-Boltzmann law* and it is as follows:

$$M_0 = \sigma T^4 \ ,$$

where M_0 is the power emitted by the blackbody per unit area and $\sigma = 5.67{\cdot}10^{-8}$ $\mathrm{Wm}^{-2}\mathrm{K}^{-4}$ is the Stefan-Boltzmann constant. Using the Stefan-Boltzmann law, Willy Wien [1896] succeeds in deriving an expression for the blackbody emission spectrum:

$$J(\lambda, T) = \frac{c_1}{\lambda^5} e^{-\frac{c_2}{\lambda T}} \ ,$$

where c_1 and c_2 are constants. This spectral energy distribution by Wien agreed with the measurements available at the time. Later experiments would, however, disagree with the Wien distribution and prove to be inexplicable by classical Newtonian mechanics (therefore leading to the quantum theories discussed in Section 2.5).

With respect to quantitative theories of light scattering, John William Strutt [1871], who was later the third Baron Rayleigh, is able to explain the colours of the sky using very simple arguments involving scattering of light waves. Assuming (as others before him) that the atmosphere has a suspension of particles which are very small compared to all the visible wavelengths, Rayleigh finds

that for particles of this size, the ratio of the intensity of scattered to incident light varies inversely as the fourth power of the wavelength. This means that the shorter blue wavelengths are scattered more frequently in the atmosphere than the longer red wavelengths, and this is the cause of the blue sky and the red sunrises and sunsets. This type of scattering is today referred to as *Rayleigh scattering*.

A few decades later, a more general result describing the scattering of plane waves of light by spherical particles was derived by Ludvig Lorenz [1890]. Later Gustav Mie [1908] derives the same equations once again, but he uses Maxwell's electromagnetic field instead of a simpler wave equation to represent the light waves and he also provides experimental verification. This theory of light scattering which is useful for deriving the scattering properties of many different materials, is today referred to as *Lorenz-Mie theory*.

Many things have been said at this point about the nature, propagation, absorption, emission, and scattering of light, but we still lack a way to combine all these ideas. What is missing, is a theory describing the flux of radiation that would be found in some particular direction at some point in a scattering medium as the result of some incident illumination progressing through the medium. Equations for such treatment of light at a macroscopic, quantitative level were introduced by Arthur Schuster [1905] in order that he could take scattering into account when considering an atmosphere. Similar equations were given a more elegant formulation by Karl Schwarzschild [1906] in his investigations of the atmosphere of the sun. The equations are a mathematical model describing the phenomenon of scattering rather than they are based on physical foundations such as Maxwell's equations. But they have subsequently been shown to give correct results in most cases. The mathematical formulations of Schuster and Schwarzschild became the birth of the quantitative *radiative transfer theory*.

During the following years the radiative transfer theory developed to a very advanced state with analytical solutions for many special cases. With a series of papers on multiple scattering Subrahmanyan Chandrasekhar was an important influence in this development. In 1950 he published the first definitive text of the field [Chandrasekhar 1950] and it is still today an important reference in all works on the subject.

An abundance of the results that have been discussed in this section are used extensively in computer graphics. The quantitative theories are particularly useful because we have to compute the visual effects due to light scattering in complicated scenarios. It is in other words of great importance that the theories we use work at a macroscopic level. Nevertheless, we see again and again that the wave theory of light must be taken into account for the correct simulation of some visual phenomenon.

2.3 Realistic Image Synthesis

In the early days of computer science, an image drawn on a screen using a computer was a small miracle. The first computer graphics were created in 1950 by Ben F. Laposky [1953]. He generated artistic works using a cathode ray oscilloscope controlled by an electronic machine. Shortly after (on 20 April 1951) the MIT Whirlwind Computer was demonstrated for the first time. It had a large modified oscilloscope as its screen and was able to display text and graphics in real-time. At this time only very few people actually had access to a computer, but as the technology developed, it was used increasingly for different tasks of computation in companies. The General Motors Research Laboratories used computers for engineering and scientific analyses in 1952, and in 1959 they started implementing a system for Design Augmented by Computers (DAC-1). As part of DAC-1 they started developing hardware for graphical man-machine communication in cooperation with IBM [Krull 1994]. These commercial interests in visualisation of three-dimensional design must have made computer graphics research very appealing to young computer science researchers.

At the beginning of the 1960s, research in 3D graphics gathered momentum. Ivan E. Sutherland [1963] presents a system called Sketchpad, which allows the user to draw line drawings on a computer screen interactively using a light pen. His work is extended by Timothy E. Johnson [1963] and Lawrence G. Roberts [1963] who start developing algorithms for displaying line sketches of 3D solid shapes. This way of doing computer graphics is remarkably similar to Democritus' way of explaining vision. To simulate the appearance of an object, we print its outline onto the screen by steering the electron beam of a CRT (Cathode Ray Tube) display monitor. This resembles Democritus' idea that the object would be imprinted in the air reaching the eye.

The early computer displays were modified oscilloscopes and they displayed vector graphics, but relatively quickly the raster techniques known from TV technology became the display technology of choice. This means that the CRT monitor displays an array of dots (or picture elements - pixels) of different intensities. With this development a raster technique for line drawing was needed and in the years after it had been presented by Jack E. Bresenham [1965], the work of Johnson [1963] and Roberts [1963] was recast to suit the rasterization approach.

Based on the methods for rendering of solids as line drawings, Arthur Appel [1968] and John E. Warnock [1969] take the next logical step when they introduce shaded display of solids. Just as in the optical theories known at the time of Aristotle, Appel introduces a *ray casting* method which, in essence, corresponds to rays of sight moving in straight lines from the eye to a surface.

Warnock uses a non-physical version of Lambert's cosine law to shade each polygon differently depending on its distance and orientation towards the light source. Henri Gouraud [1971] notices the "cartoon-like" appearance of these flat shading schemes and introduces continuous shading across the polygons. This is today referred to as *Gouraud shading*, but Lambert actually also investigated this subject in the 18th century when he enunciated his cosine law [Lambert 1760]. What the graphics researchers are investigating at this point is the shading of objects which was also noticed by Ptolemy. The Lambertian-like shading introduced by Warnock captures diffuse objects, but a quantitative model is missing for rendering of specular highlights. Such a model is provided by Bui Tuong Phong [1975]. It is not physically accurate, but it is very efficient and the *Phong model* is famous in graphics today.

Following the Phong model, more physically-based reflection models for rough surfaces were imported to graphics by James F. Blinn [1977]. He even includes the Fresnel equations in his models. But the real breakthrough towards realism in computer-generated images comes when Turner Whitted [1980] introduces *ray tracing*. In analogy with Euclidean optics, rays proceed from the eye (the image plane) and interact with the surfaces they arrive at. When they interact with a diffuse surface, a shading model is used. When they interact with a specular surface, the laws of reflection and refraction are employed and new rays are traced in the specular directions.

Based on ray tracing, several techniques related to the ancient mathematical optics were tested in the first half of the 1980s. Cone tracing [Amanatides 1984], for example, is closely related to the Ptolemaic concept of radiation distributed continuously in the visual cone. Many of the results discovered by both Euclid, Diocles, and Ptolemy in their works on optics and catoptrics, are useful in cone tracing. Distribution ray tracing [Cook et al. 1984] takes into account how radiation is spread at each point of intersection. This corresponds to the spread of radiation investigated by Ptolemy and al-Kindī. A few years later, graphics also find a use for the idea of letting rays issue from the light sources instead of the eyes [Arvo 1986]. At the time it is called "backward ray tracing" as the tracing direction is opposite to the usual approach. This terminology is later abandoned because it easily causes confusion. *Light ray tracing*, as we now call it, allows us to capture light phenomena known as *caustics* more easily. Caustics are the bright highlights that occur when light has been focussed through a lens or a fluid onto a surface. In a way this is similar to al-Haytham and Kepler's discovery that the eye is merely a lens focussing light on the retina. Fermat's principle was introduced for rendering of mirages [Berger and Trout 1990] at a rather late time in graphics.

Being far ahead of his time, Blinn [1982] introduces scattering effects inspired by thermal radiation and the work of Chandrasekhar [1950], and only a few

years later the subject of realistic rendering is firmly connected to heat transfer by Nishita and Nakamae [1983; 1985] and Goral et al. [1984], and to radiative transfer theories by James T. Kajiya [1984]. The methods inspired by heat transfer are referred to as *radiosity* methods in graphics. They capture the colour bleeding effects observed by Galen. The rendering methods based on radiative transfer are the most general methods used in graphics so far.

Towards the beginning of the 1990s the discipline of realistic rendering starts proceeding in two directions: One branch seeking algorithms that are fast enough to allow real-time interaction with rendered scenes, and another branch seeking improved realism without worrying about the time it takes to render it. In the latter branch graphics has continued to move closer to the wave theories of light. The connection between trichromatic (Young-Helmholtz) colour theory and wavelengths has been introduced in graphics [Meyer and Greenberg 1980], and Bouguer-Lambert's law and the Fresnel equations have been incorporated as standard elements in realistic rendering [Glassner 1995]. A rendering method based on simplified wave theory is first considered for graphics by Hans P. Moravec [1981]. This approach is, however, very expensive and subsequent methods based on wave theory have mostly been used to derive local shading models [Kajiya 1985; Bahar and Chakrabarti 1987]. In a slightly different order than the development described in Section 2.2 (rather following the difficulties in implementation), but not completely off target, we have seen graphics simulations of dispersion [Thomas 1986], interference [Smits and Meyer 1990; Dias 1991], birefringence and polarisation [Tannenbaum et al. 1994], and diffraction effects [Stam 1999]. Even blackbody emission is used in graphics to compute emission spectra for natural light sources [Stam and Fiume 1995; Glassner 1995] and Rayleigh scattering is used for rendering of a realistic sky [Klassen 1987].

More recently the Lorenz-Mie theory has been used [Rushmeier 1995; Callet 1996] for computing the coefficients needed in the realistic rendering methods that are based on radiative transfer theory. This means that we can find macroscopic input coefficients using Maxwell's equations, but we are yet to see a complete rendering method based on the electromagnetic field theory. Of the rendering methods currently in use, the ones closest to the wave theories of light are the ones based on geometrical optics. This type of rendering was introduced by Stam and Languénou [1996]. Geometrical optics is a simplification of Maxwell's equations which assumes that the wavelength of light is so small that we can think of light as rays following trajectories which are not necessarily straight. In other words, the methods currently in use are simply the correct way to handle heterogeneous media using ray tracing. The attempt by Moravec [1981] is to my knowledge still the only attempt on a complete rendering method based on the wave theory of light.

The real-time branch is different in the sense that it more often chooses a com-

promise between physics and a simplified mathematical model. The Phong model is a good example. It is often the case that old theories, like the ones we have discussed throughout this chapter, present a simple, but not entirely physically correct explanation of a light phenomenon. These old mathematical models often give surprisingly good visual results and their simplicity makes them well suited for real-time implementation on modern programmable hardware. This means that we can still find useful mathematical models for computer graphics by digging into the history of the theories of light. As an example it is shown, in the following section, how Aristotle's rainbow theory makes us able to render rainbows in real-time. Rendering rainbows more physically correctly is certainly not a real-time process. It has been done by Jackèl and Walter [1997] as follows. Lorenz-Mie theory is used to compute how small water drops scatter light. The result is used as input for the radiative transfer equation, which is used for a full volume visualisation of the air containing the water drops. This expensive computation captures the rainbow. Let us see how the Aristotelian theory works.

2.4 Rendering the Aristotelian Rainbow

Aristotle's [~350 B.C., 1984] theory for the formation of rainbows is really quite simple. We have discussed it briefly in Section 2.1, but now we will add a few extra details. Aristotle thinks of the sky as a hemisphere (see Figure 2.2). This is very similar to the sky domes used in graphics. We draw a sphere with inward facing polygons and map a texture onto it to obtain a sky rendering. If we place a sun on the sky dome, it makes us able to use Aristotle's rainbow theory.

For every point on the dome that we render, we will know the direction toward the eye $\vec{\omega}$ and the direction toward the sun $\vec{\omega}'$. Aristotle's theory states that the rainbow forms on the hemisphere of the sky where the angle between $\vec{\omega}$ and $\vec{\omega}'$ is equal. Aristotle does not say what the angle is, however from newer rainbow theories we know that 42° is a good choice. If the sun were a point source, the result would be an infinitely thin circular arc reflecting the intensity of the sun when

$$\vec{\omega} \cdot \vec{\omega}' = \cos 42^\circ \approx 0.7431 \ .$$

This test is easily done in a fragment shader on modern graphics hardware. To get a rainbow instead of a bright line across the sky, we simply take into account that the sun has an extension which covers a range of directions $\vec{\omega}'$. We use the lowest and the highest point of the sun on the sky dome. This gives two cosine values

$$a = \vec{\omega} \cdot \vec{\omega}'_{\text{high}} \quad , \quad b = \vec{\omega} \cdot \vec{\omega}'_{\text{low}} \ .$$

With a `smoothstep` function (Hermite interpolation between a and b) we use a and b to find a shade value $c \in [0, 1]$ for the rainbow:

$$c = \texttt{smoothstep}(a, b, 0.7431) \ .$$

To let the value c determine the colour of the rainbow, we have to involve some more recent colour theory. Each value of c corresponds to a wavelength in the visible spectrum such that $\lambda(c = 0) = 400\,\text{nm}$ and $\lambda(c = 1) = 780\,\text{nm}$. To find RGB colour values for each wavelength, we use the RGB colour matching functions [Stiles and Burche 1959; Stockman and Sharpe 2000]. A look-up using c in a 1D texture holding the colours of the rainbow, i.e. the RGB colour matching functions, is one way to get the desired colours. Another option is to choose a few RGB colours at significant wavelengths (e.g. at $\lambda = 445\,\text{nm}, 540\,\text{nm}, 600\,\text{nm}$) and then interpolate between them using c. Finally an alpha value is used to blend the rainbow with the background sky. For $c = 0$ and $c = 1$ the alpha value is 0 (the rainbow does not show in these regions), in-between that the alpha value should be set depending on how intensely the user wants the rainbow to appear in the sky.

The Aristotelian rainbow is very simple to render and it is easily run in real-time. It runs at 116 frames per second in a 1200×400 resolution on an NVIDIA GeForce Go 7400 graphics card. Sample renderings are shown in Figure 2.1 (page 16) and in Figure 2.3 (page 36). Figure 2.3 also shows where the lowest and highest points of the sun are placed in the sky. The distance between these two points determine the width of the rainbow. In addition, it is easy to modify the position and intensity of the rainbow in the sky by moving the points and adjusting the alpha value. Originally Aristotle thought of the sun as sitting on the hemisphere (the sky dome), where the rainbow also appears. This gives rainbows which are very stretched out compared to real rainbows. If we move the points on the sun, which determine $\vec{\omega}'_{\text{high}}$ and $\vec{\omega}'_{\text{low}}$, away from the sky dome in the radial direction, the rainbow gets a more natural arc. This is another parameter we can use to modify the appearance of the Aristotelian rainbow. Using these different parameters, we have, qualitatively, tried to match the appearance of real rainbows in Figure 2.4 (page 37).

With the Aristotelian rainbow we have given a brief example of what we can learn by taking an interest in the history of theories of light and vision. The first sections of this chapter present pointers to relevant developments in the history of these theories. Hopefully these references provide an overview and a starting point for finding more mathematical and physical models that can be useful in graphics. In the next section, we follow the development of the theories of light further to get ideas and inspiration about possible future developments in realistic image synthesis.

Figure 2.3: *Rainbows rendered in real-time using Aristotle's theory for rainbow formation. The black and green points (in the smaller images) show the lowest and the highest point of the sun. They determine the position and size of the rainbows (in the larger images). Of course the rainbow is always found when we look in the direction opposite the sun.*

Figure 2.4: *Comparison of rainbow pictures from the real world (top row) and rainbows rendered using Aristotle's theory (bottom row).*

2.5 Quantum Theories

Not long after Wien [1896] proclaimed his functional expression for the blackbody emission spectrum, Rubens and Kurlbaum [1900] carry out new measurements that show a definite disagreement with the Wien distribution at longer wavelengths. Numerous new expressions are then proposed to match these measurements, but the simplest fit is found by Max Planck [1900a]. His fit is

$$L(\lambda, T) = \frac{c_1 \lambda^{-5}}{e^{\frac{c_2}{\lambda T}} - 1} \ .$$

Having found this simple and surprisingly precise fit, Planck devotes himself to providing a physical justification for it. His theory is published a few months later [Planck 1900b]. As a part of his theory he introduces the universal constant h which is now called *Planck's constant* (its modern value is $h = 6.63 \cdot 10^{-34}$ Js). He writes [Planck 1900b, p. 239]:

> We consider then - and this is the most essential point of the entire calculation - [the energy] E to be a compound of a definite number of finite equal parts and we accommodate this by using the natural constant h [.]

Planck refers to these "finite equal parts" as energy elements given by $\varepsilon = h\nu$, where $\nu = c/\lambda$ is the frequency of the light. This is the birth of the quantum theories of light. With this assumption Planck is able to describe the constants c_1 and c_2 in his blackbody emission spectrum using the universal constants h, k, and c (where the latter two are the Boltzmann constant and the speed of light in a vacuum). He finds $c_1 = 8\pi ch$ and $c_2 = hc/k$.

Blackbody radiation is not the only light phenomenon which turns out to be inexplicable (quantitatively) by Maxwell's electromagnetic field theory. Another phenomenon is the production of electric current using light. This is first observed in 1839 by Alexandre Edmund Becquerel [1868, p. 122] and today we call it the *photoelectric effect*. By an experiment which shows that ultraviolet light is able to cause an electric discharge, Hertz [1887] starts off a more thorough investigation of the subject. The existence of the photoelectric effect is perhaps not surprising considering that light is electromagnetic waves, but matching the observed quantity of electrons liberated by a beam of light, turns out to be extremely troublesome. After more than a decade with no satisfactory explanation, Philipp Lenard [1902] finds experimental evidence that the photoelectric effect can hardly agree with the electromagnetic field theory.

Inspired by Planck's blackbody emission spectrum and the experiments of Lenard, Albert Einstein [1905] presents a theory which radically departs from the

wave theories of light. He proposes that Planck's energy elements exist as real particles of light or "light quanta". Einstein's light quanta are today known as *photons*. With this theory Einstein is able to give a convincing explanation of the photoelectric effect. Further extending his work, Einstein [1906] uses Planck's blackbody radiation theory to show how light must be emitted and absorbed in jumps which are integral multiples of $h\nu$.

While Planck supported most of Einstein's work, he did not accept the idea of light quanta until many years later [Mehra and Rechenberg 1999, fn. 50]. Trying to take a step away from the quantum theory and closer to the classical theories, Planck [1912] proposes a second theory for the derivation of his blackbody emission spectrum. In this theory he assumes that energy is emitted in discrete quanta, but absorbed continuously. Employing statistical theory to determine the probability that a material (an oscillator to be precise) has absorbed enough energy to emit an energy quantum, Planck arrives at a curious formula for the internal energy U of the material (oscillator) which is as follows [Planck 1912, p. 653]:

$$U(\nu, T) = \frac{h\nu}{2} \frac{e^{\frac{h\nu}{kT}} + 1}{e^{\frac{h\nu}{kT}} - 1} = \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1} + \frac{1}{2} h\nu \ .$$

Observe that for $T \to 0$, $U \to \frac{1}{2} h\nu$. This means that a material at temperature absolute zero still has internal energy. Planck did not emphasise this, he merely took note of it, but it was the birth of the concept of *zero-point energy* [Milonni and Shih 1991].

To find a reason for Planck's concept that energy is emitted from an atomic system in distinctly separated quanta, Niels Bohr [1913] proposes a model of the hydrogen atom. His model explains how an atom has a discrete set of energy states such that it will emit energy at frequency $\nu = (E_m - E_n)/h$ when passing from one energy state E_m to another E_n. This is sometimes referred to as *Bohr's frequency condition*. Bohr's model provided exactly what Einstein was missing in order to find a derivation of Planck's spectrum based on his concept of light quanta. For his derivation Einstein [1916] introduces the concepts of *spontaneous* and *stimulated* emission. Let A_{mn} and B_{mn} denote the rate of spontaneous and stimulated emission respectively, then the distribution of blackbody radiation becomes [Einstein 1916, p. 53]:

$$\rho(\lambda, T) = \frac{A_{mn}/B_{mn}}{e^{\frac{E_m - E_n}{kT}} - 1} = \frac{A_{mn}/B_{mn}}{e^{\frac{h\nu}{kT}} - 1} \ .$$

And after a short argument about A_{mn}/B_{mn} (using Wien's displacement law and the classical limit for high temperatures), Einstein finds that this distribution agrees with Planck's spectrum. As a part of his theory Einstein deduces that light quanta must carry momentum. In a way this is the prediction of the existence of light particles and it is confirmed by Arthur Compton who writes

that "remarkable *agreement between experiment and theory* indicates clearly [...] that a radiation quantum carries with it momentum as well as energy" [Compton 1923, p. 484].

Einstein did not include zero-point energy in his 1916 derivation of the Planck spectrum. Of course, he did not have to include it as there had yet been given no direct experimental evidence for its existence. Around a decade later such evidence does, however, appear. Mulliken [1925] shows the existence of half-integral quantum numbers and James and Firth [1927] give the first direct proof of zero-point motion. Half a year after Mulliken's observation of half-integral quantum numbers, Werner Heisenberg [1925] presents the theoretical justification for their existence as well as for the existence of zero-point energy. He does this by deriving that the energy of a Planck oscillator is $E = (n + \frac{1}{2})h\nu$ [Heisenberg 1925, p. 889]. The derivation is done as part of an example to show the applicability of quantum mechanics which he lays out the foundations for in the very same paper.

With these new developments, Paul Adrien Maurice Dirac [1927] is able to introduce the theory of *quantum electrodynamics*, which is the theory that concerns the interaction of photons and electrically charged particles (electrons and positrons). This is a theory which "leads to the correct expressions for Einstein's A's and B's" [Dirac 1927, p. 265]. The correct A and B coefficients make us able to derive the correct blackbody emission spectrum. From the standpoint of quantum electrodynamics it is [Milonni and Shih 1991, p. 688]

$$L(\lambda, T) = \frac{8\pi ch\lambda^{-5}}{e^{\frac{hc}{\lambda kT}} - 1} + 4\pi ch\lambda^{-5} \ .$$

The theory of quantum electrodynamics has been refined and improved many times after its introduction and it has been tested over an extremely wide range of conditions. Still no significant disagreements have been found between experiment and theory. In the words of Richard P. Feynman [1985, pp. 7–8] "the theory describes *all* the phenomena of the physical world except the gravitational effect [...] and radioactive phenomena, which involve nuclei shifting in their energy levels". Quantum electrodynamics is, in other words, the most exact theory of light that we have.

Why do we not use quantum electrodynamics in realistic image synthesis? The simple answer is that it is too complicated. The phenomena that quantum electrodynamics describe happen at a scale which is so microscopic that any scenario which is not an extremely simplified laboratory setup, is just not described in a feasible way using this theory. Why even bother to write about quantum electrodynamics then? For three reasons:

1. It is the key to understanding the interaction of light and matter. If we want an exact theory for computing macroscopic material properties using the microscopic structure of a material, then the theory of quantum electrodynamics will eventually be indispensable.

2. If we start from the most exact theory and keep simplifying it until we obtain the models that we use in practice, then we will know exactly what phenomena our models are able to simulate and what they are not. And we will know what simplifying assumptions we need to work on if we want our models to capture more phenomena.

3. The enormous complexity of cases considered in quantum electrodynamics have led researchers to important general principles for handling problems of high dimensionality. If we look into some of the acute solutions that have been presented over the years, we might stumble upon mathematical tools that are of high relevance toward realistic rendering.

The surface has been scratched by Andrew Glassner who investigated the first point (1.) by describing the quantum-mechanical structure of materials in Chapter 14 of his monumental work on the principles of digital image synthesis [Glassner 1995]. Very recently Banks and Abu-Raddad [2007] did the first work concerning the second point (2.) by going from quantum electrodynamics to Maxwell's equations to describe the foundations of photo-realistic rendering. To further investigate these three reasons for considering quantum electrodynamics in graphics, the remaining chapters of Part I are about finding the limiting assumptions that we rely on when we go from quantum electrodynamics to mathematical models that are practical for realistic rendering (2.). Part II is about matter and is consequently related to the first point (1.). Finally, and perhaps surprisingly, some introductory remarks are made about the third point (3.) in Part III.

Chapter 3

Quantum Electrodynamics

One of the most extraordinary and exciting things about modern physics is the way the microscopic world of quantum mechanics challenges our common-sense understanding.

His Holiness the Dalai Lama, from *The Universe in a Single Atom*

This chapter serves four purposes (a-d). The first purpose (a) is to explain, introductorily, the properties of quantum particles. The word "photon" is used in many contexts in graphics, but it is hardly ever used to denote the quantum particle to which the name belongs. This sometimes causes confusion among graphics students as to what a photon actually is. The idea in explaining quantum particles is therefore both to introduce the subject of this chapter and to shed some light on the true meaning of the word "photon". The second purpose (b) is to connect quantum electrodynamics to Maxwell's equations, and, in doing so, to extract the simplifying assumptions that we make when we adopt a wave theory of light. This purpose is very similar to the purpose of the paper by Banks and Abu-Raddad [2007]. The approach in this chapter is, however, very different from their approach. The third purpose (c) is to introduce the concept of operators working on systems with many degrees of freedom. The creation and annihilation operators are essential in quantum electrodynamics. When they are introduced in the following, some remarks will be given on the general principles behind the operators. The fourth purpose (d) is to outline how one could construct a rendering algorithm based on the principles of quantum electrodynamics. This outline is given as concluding remarks in Section 3.4.

It is convenient to divide an introduction to quantum electrodynamics into three parts: one to describe photons moving through free space (Sec. 3.1), one to describe electrons moving through free space (Sec. 3.2), and one to describe the interaction of photons and electrons (Sec. 3.3). This reflects "the three basic actions, from which all the phenomena of light and electrons arise" [Feynman 1985, p. 85]. These are described by Feynman [1985, p. 85] as follows:

- ACTION 1: A photon goes from place to place.
- ACTION 2: An electron goes from place to place.
- ACTION 3: An electron emits or absorbs a photon.

The three basic actions are governed by the total energy of a system of quantum particles (because the function, or operator, which describes the total energy gives rise to the equations of motion [von Neumann 1955]). In this chapter we describe the total energy in terms of a Hamiltonian operator. A Hamiltonian operator will therefore be described for each of the three basic actions in each of the three sections to follow.

To understand quantum electrodynamics, we first have to get an idea of the concept of quantum particles. This is not so easily accomplished. We cannot think of photons as balls flying around in straight lines and bouncing off surfaces. In the words of Feynman [1963, Sec. 37-1]:

> Things on a very small scale behave like nothing that you have any direct experience about. They do not behave like waves, they do not behave like particles, they do not behave like clouds, or billiard balls, or weights on springs, or like anything that you have ever seen.
>
> [...]
>
> We know how large objects will act, but things on a small scale just do not act that way. So we have to learn about them in a sort of abstract or imaginative fashion and not by connection with our direct experience.

The first thing we need to know about quantum particles is that it is *fundamentally impossible* to determine the exact location x and momentum p of a particle. Suppose we want to determine the position of a particle with uncertainty Δx and the momentum with uncertainty Δp, then *Heisenberg's uncertainty principle* says that $\Delta x \, \Delta p \geq \hbar/2$, where $\hbar = h/(2\pi)$. All of quantum mechanics rely on the validity of this principle and so far it has never been proven wrong. The impact of the uncertainty principle is considerable. It means that we cannot say exactly what will happen. All we can determine is the probability of some

event to occur. Consequently, we have to think differently when working with quantum theories as compared to when we use the classical laws of physics.

Suppose we think of a *system* as a setup where all initial and final conditions are completely specified. Then following Feynman, we define an *event* to be a specific set of initial and final conditions. In a system we let P denote the probability of an event. With every event we associate a complex number ϕ which is referred to as the *probability amplitude* of that event. A system of quantum particles has the following properties [Feynman et al. 1963, Sec. 37-7]:

1. $P = |\phi|^2$.
2. If an event can occur in several different ways, the probability amplitude for the event is the sum of the amplitudes for each separate way $\phi = \phi_1 + \phi_2 + \dots$.
3. If we are able to determine whether one or another alternative is taken, the probability of the event is the sum of the probabilities for each alternative $P = P_1 + P_2 + \dots$.

This means that probability amplitudes of indistinguishable ways in which an event can occur interfere like waves. However, "one cannot design equipment in any way to determine which of two alternatives is taken, without, at the same time, destroying the pattern of interference" [Feynman et al. 1963, Sec. 37-8]. According to Feynman et al. [1963] this is a more general statement of the uncertainty principle. It is another way of saying that we are unable to predict precisely what a particle will do.

Any system of quantum particles can be separated by a filtering process into a certain set of *base states*. The base states are independent in the sense that the future behaviour of particles in any given base state depends only on the nature of that particular base state [Feynman et al. 1965, Sec. 5-4]. Any particle has a number of base states. The *state of a particle* is a set of probability amplitudes which contains the amplitudes for the particle to be in each of its base states. Consider a particle in state s_1. The probability amplitude that the particle will end up in state s_2 is denoted $\langle s_2|s_1\rangle$. If we construct a vector basis using the base states of a system (for instance a system which describes the initial and final conditions of a single particle), we have means to obtain a *state vector* for every possible state of the system (particle).

Using Dirac's "bra-ket" notation [Dirac 1930], a "to" state vector is called a *bra*, and is denoted $\langle s_2|$, while a "from" state vector is called a *ket*, and is

denoted $|s_1\rangle$. Thinking of i as any one of the base states, it follows that [Feynman et al. 1965, Sec. 8-1]

$$\langle s_2|s_1\rangle = \sum_i \langle s_2|i\rangle\langle i|s_1\rangle \ . \tag{3.1}$$

This formula reveals how every possible state vector is a linear combination of the base state vectors:

$$|s\rangle = \sum_i |i\rangle\langle i|s\rangle \ . \tag{3.2}$$

For probability to be "conserved" in a system, it is generally true that [Feynman et al. 1965, Sec. 5-5]

$$\langle s_2|s_1\rangle = \langle s_1|s_2\rangle^* \ , \tag{3.3}$$

where the asterisk * denotes the complex conjugate.

A common choice of base states for a quantum particle comprises the states of definite momentum and the states of angular momentum along some axis. A particle can have infinitely many different definite momenta. This turns the summations (3.1,3.2) into integrations. A peculiar property of quantum particles is that their angular momentum is always an integer or a half-integer. The *spin* of a particle determines the range of integers or halves that the angular momentum can attain. If the particle has spin j, its angular momentum along any particular axis will have one of the values [Feynman et al. 1964, Sec. 35-1]

$$-j\hbar, (-j+1)\hbar, \ldots, (j-1)\hbar, j\hbar \ ,$$

According to this rule, the angular momentum of a spin one particle gives rise to a three-fold infinite set of base states (for each definite momentum there are three amplitudes, call them $-, 0, +$). The photon is a spin one particle, but it *always* move at the speed of light and therefore cannot exist in the rest (0) state. Photon spin gives rise to the polarisation of a beam of light. Since photons have only two possible angular momenta $(-, +)$, we are able to capture all types of polarisation using two independent components (sometimes referred to as the $\perp$-polarised and the $\parallel$-polarised components).

3.1 The Free Electromagnetic Field

Consider a particle in some state ψ. It is a general principle in quantum mechanics that the probability $\langle \boldsymbol{x}|\psi\rangle$ for a particle to be found precisely at the coordinate $\boldsymbol{x}$ is proportional to $e^{+(i/\hbar)\,\boldsymbol{p}\cdot\boldsymbol{x}}$, where $\boldsymbol{p}$ is the definite momentum of the particle [Feynman et al. 1965, Sec. 16-2]. This shows that we can equally well choose the position coordinates $\boldsymbol{x}$ (along with the angular momenta) as base states for the particle. Since particle states in general are functions of

time, the same principle can help us realise that the probability amplitude for the particle to be at $\boldsymbol{x}$ follows a time-varying wave function:

$$\langle \boldsymbol{x}|\psi(t)\rangle = \psi(\boldsymbol{x},t) = Ae^{-(i/\hbar)(E_p t - \boldsymbol{p}\cdot\boldsymbol{x})} \ .$$

By comparison to a classical wave function $Ae^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})}$, this equation reveals two fundamental properties of quantum particles, namely that particle energy E_p is related to the angular frequency $\omega = 2\pi\nu$ of a wave and that definite momentum is related to the wave number $\boldsymbol{k}$:

$$E_p = \hbar\omega \quad , \quad \boldsymbol{p} = \hbar\boldsymbol{k} \ .$$

In other words, we can describe a quantum particle as a harmonic oscillator. If we use position coordinates instead of definite momenta in the set of base states for the particle (such that we have $\langle \boldsymbol{x}|\psi(t)\rangle = \psi(\boldsymbol{x},t)$), we call it the coordinate representation.

The development of a state vector $|\psi(t)\rangle$ is governed by the Schrödinger equation (developed by Erwin Schrödinger [1926] in a less general form):

$$i\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle \ . \tag{3.4}$$

In this equation $\hat{H}$ is an operator describing the total energy E of the system such that $\hat{H}|\psi\rangle = E|\psi\rangle$. It is called the *Hamiltonian operator*.

Now we should recall from Bohr's [1913] atomic model (cf. Section 2.5) that a particle can only be in certain definite energy states E_n. These are the characteristic states, or the *eigenstates*, of the Hamiltonian operator. Using this fact, we can write the probability amplitude $\psi(\boldsymbol{x},t) = \langle \boldsymbol{x}|\psi(t)\rangle$ as a sum of probability amplitudes

$$\langle \boldsymbol{x}|\psi(t)\rangle = \sum_n \langle \boldsymbol{x}|E_n\rangle\langle E_n|\psi(t)\rangle \ .$$

Each term in the sum denotes the amplitude for the particle in state ψ at time t to be in energy state E_n and at position $\boldsymbol{x}$. In a different notation we have [Milonni 1994]

$$\psi(\boldsymbol{x},t) = \sum_n \psi_n(\boldsymbol{x})a_n(t) \ , \tag{3.5}$$

where $a_n(t) = a_n(0)e^{-iE_n t}$. To interpret a_n and ψ_n, we see that a_n simply tells us the amplitude for the particle in state ψ at time t to be of energy E_n, while ψ_n tells us the amplitude for the particle to be at location $\boldsymbol{x}$. If we take the complex conjugate (3.3) we have a particle at $\boldsymbol{x}$ and find the amplitude for it to be in energy state E_n and the amplitude for that to be in the state ψ:

$$\psi^*(\boldsymbol{x},t) = (\langle \boldsymbol{x}|E_n\rangle\langle E_n|\psi(t)\rangle)^* = \langle E_n|\psi(t)\rangle^*\langle \boldsymbol{x}|E_n\rangle^* = \langle\psi(t)|E_n\rangle\langle E_n|\boldsymbol{x}\rangle \ .$$

Notice how the complex conjugate a_n^* means that we know the particle is in energy state E_n, while the original a_n means that there is an amplitude that the particle is of energy E_n. When working with more than one particle we will exploit this feature of a_n.

The states we have discussed so far have all been single-particle states. The *state of a system* is also described by a state vector $|\Psi\rangle$. It is similar to the state of a particle, only it combines base states for all the particles involved. We can think of the system as an outer product of the configuration spaces of all the involved particles. This means that a state vector typically has a many-fold infinite number of dimensions. One way to cope with this immense number of degrees of freedom is to use operators. In the standard formalism of quantum electrodynamics, creation and annihilation operators are introduced. Let us try to follow the standard formalism.

Consider a system of several particles. Since only an integral number of particles can exist, we *quantize* the system. Quantization is done by changing the meaning of a_n. Instead of using the amplitude a_n for a single particle to be of energy E_n, we use an operator $\hat{a}_n$ which *annihilates* a particle of energy E_n from the system state vector. In analogy with the complex conjugate a_n^*, we let $\hat{a}_n^\dagger$ denote the (Hermitian) conjugate operator which *creates* a particle of energy E_n. When we use the creation and annihilation operators on a state $|N_1, \ldots, N_m\rangle$ involving N_n particles of energy state E_n $(n = 1, \ldots, m)$, we get the results:

$$\begin{aligned}
\hat{a}_n|N_1, \ldots, N_m\rangle &= \sqrt{N_n}\,|N_1, \ldots, N_n - 1, \ldots, N_m\rangle \\
\hat{a}_n^\dagger|N_1, \ldots, N_m\rangle &= \sqrt{N_n + 1}\,|N_1, \ldots, N_n + 1, \ldots, N_m\rangle \\
\hat{a}_n^\dagger\hat{a}_n &= N_n \\
\hat{a}_n\hat{a}_n^\dagger &= N_n + 1\ .
\end{aligned}$$

Note that $\hat{a}_n$ and $\hat{a}_n^\dagger$ do not commute. In terms of the many-fold infinite space which the state vectors live in, we can think of the creation operator $\hat{a}_n^\dagger$ as an outer product adding the configuration space of yet another particle to the system, and we can think of the annihilation operator $\hat{a}_n$ as a projection removing the configuration space of a particle. Thus the creation and annihilation operators comprise the fundamental principles of expansion and reduction of dimensionality to handle a system with many degrees of freedom (we will return to this point in Chapter 14).

From the quantum point of view a field of photons in free space must have a total energy given by a sum over the energies of the photons in the field. And since there is experimental evidence (cf. Section 2.5) for the existence of zero-point energy, we should include that in the sum. The total energy of a photon

field in free space is then

$$E = \sum_{n=1}^{m} E_n \left(N_n + \frac{1}{2} \right) = \sum_{n=1}^{m} \hbar\omega_n \left(N_n + \frac{1}{2} \right) \ ,$$

where m is the number of photon energy states in the field and ω_n is the angular frequency of a photon in state n. As the Hamiltonian operator obtains the total energy from a state vector, we can describe the Hamiltonian operator of a photon field in free space using the creation and annihilation operators:

$$\hat{H} = \sum_{n=1}^{m} \hbar\omega_n \left(\hat{a}_n^\dagger \hat{a}_n + \frac{1}{2} \right) = \frac{1}{2} \sum_{n=1}^{m} \hbar\omega_n (\hat{a}_n \hat{a}_n^\dagger + \hat{a}_n^\dagger \hat{a}_n) \ .$$

If we express the creation and annihilation operators in terms of position or definite momentum, and insert the Hamiltonian operator in the Schrödinger equation (3.4), we have the equation governing the first of the three basic actions ("a photon goes from place to place"). Let us first use definite momenta.

Suppose we let the photon energy states (E_n) be determined by definite momentum $\boldsymbol{p}$ and angular momentum σ of the particle, then we denote the index of an energy state $\boldsymbol{k}\sigma$, where we use the relationship between definite momentum and the wave number $\boldsymbol{p} = \hbar\boldsymbol{k}$. If we assume a continuous distribution of definite momenta, we have

$$\hat{H} = \sum_{\sigma=-,+} \frac{1}{2(2\pi)^3} \int \hbar\omega_{\boldsymbol{k}\sigma} (\hat{a}_{\boldsymbol{k}\sigma} \hat{a}_{\boldsymbol{k}\sigma}^\dagger + \hat{a}_{\boldsymbol{k}\sigma}^\dagger \hat{a}_{\boldsymbol{k}\sigma}) \, \mathrm{d}\boldsymbol{k} \ . \tag{3.6}$$

This assumption is only valid when the number of photons is large (and that is usually the case). Having found an expression to govern the first of the three basic actions, the next step is to see if we can connect it to the theory of classical electromagnetic fields.

The total energy E_M (M for Maxwell) of a classical electromagnetic field in free space is the integral over the energy density of the field at all positions in space [Martin and Rothen 2004]:

$$E_M = \frac{\varepsilon_0}{2} \int (|\boldsymbol{E}|^2 + c^2 |\boldsymbol{B}|^2) \, \mathrm{d}\boldsymbol{x} \ , \tag{3.7}$$

where $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic field vectors. Our job is to investigate whether this is the same total energy as what we would obtain using the Hamiltonian operator (3.6) on the system state vector.

To do this job, we need the vector potential $\boldsymbol{A}$. It is defined indirectly in terms of the electric and magnetic field vectors [Feynman et al. 1964, Sec. 18-6]:

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \quad , \quad \boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t} \ , \tag{3.8}$$

where ϕ is the scalar potential. With this definition the vector and scalar potentials are not unique. We can change them without changing the physics of the system using the rule:

$$\boldsymbol{A}' = \boldsymbol{A} + \nabla\chi \quad , \quad \phi' = \phi - \frac{\partial\chi}{\partial t} \ ,$$

where χ is a given function of space and time. We choose a specific form of χ by choosing an equation involving $\nabla \cdot \boldsymbol{A}$ that χ must satisfy. This is called choosing a gauge. In the Coulomb gauge we have

$$\nabla \cdot \boldsymbol{A} = 0 \ .$$

In the absence of any sources (and therefore also in a free field) we have $\phi = 0$. This means that another way to write the total energy of a free electromagnetic field in the Coulomb gauge is

$$E_M = \frac{\varepsilon_0}{2} \int \left(\left| \frac{\partial \boldsymbol{A}}{\partial t} \right|^2 + c^2 |\nabla \times \boldsymbol{A}|^2 \right) \mathrm{d}\boldsymbol{x} \ . \tag{3.9}$$

where $\varepsilon_0 = 8.8542 \cdot 10^{-12}\,\mathrm{F/m}$ is the vacuum permittivity.

The Fourier expansion of the vector potential is

$$\boldsymbol{A}(\boldsymbol{x}, t) = \sum_{\sigma=-,+} \frac{1}{(2\pi)^3} \int \boldsymbol{c}_{\boldsymbol{k}\sigma}(t) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} \, \mathrm{d}\boldsymbol{k} \ . \tag{3.10}$$

Since the vector potential $\boldsymbol{A}$ is a real vector, we have $\boldsymbol{c}^*_{\boldsymbol{k}\sigma} = \boldsymbol{c}_{-\boldsymbol{k}\sigma}$ and

$$\int \boldsymbol{A}^2(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x} = \sum_{\sigma=-,+} \frac{1}{(2\pi)^3} \int \boldsymbol{c}_{\boldsymbol{k}\sigma}(t) \cdot \boldsymbol{c}_{-\boldsymbol{k}\sigma}(t) \, \mathrm{d}\boldsymbol{k} \ . \tag{3.11}$$

Equations 3.9 and 3.11 constitute an expression for the total energy of a system in terms of an integral over a continuous distribution of definite momenta. In this way the Fourier expansion of the vector potential provides the link between the classical electromagnetic field and the quantized photon field.

To make the connection between the energy E_M of a classical electromagnetic field and the energy E, which is given by $E|\Psi\rangle = \hat{H}|\Psi\rangle$, we manipulate the expression for the Hamiltonian operator (3.6). To make the notation less cluttered, we use the following operators instead of the creation and annihilation operators:

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}(t) &= \textstyle\sum_\sigma \left(\frac{\hbar}{2\varepsilon_0 \omega_{\boldsymbol{k}\sigma}} \right)^{1/2} \hat{a}_{\boldsymbol{k}\sigma}(t) \vec{e}_{\boldsymbol{k}\sigma} \\ \hat{\boldsymbol{\alpha}}^\dagger_{\boldsymbol{k}}(t) &= \textstyle\sum_\sigma \left(\frac{\hbar}{2\varepsilon_0 \omega_{\boldsymbol{k}\sigma}} \right)^{1/2} \hat{a}^\dagger_{\boldsymbol{k}\sigma}(t) \vec{e}_{\boldsymbol{k}\sigma} \ , \end{aligned} \tag{3.12}$$

where $\vec{e}_{\boldsymbol{k}\sigma}$ is a unit vector perpendicular to $\boldsymbol{k}$ in the polarisation σ. The expression for the Hamiltonian operator (3.6) is then:

$$\hat{H} = \frac{\varepsilon_0}{(2\pi)^3} \int \omega_{\boldsymbol{k}}^2 (\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}) \, \mathrm{d}\boldsymbol{k} \ . \tag{3.13}$$

The plan is to find the link between $\boldsymbol{c}_{\boldsymbol{k}\sigma}$ and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}$. The problem is that $\boldsymbol{c}_{\boldsymbol{k}\sigma}$ is real and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}\sigma}$ would be complex if we turned it back into a wave function. As a trial solution let us try

$$\hat{\boldsymbol{c}}_{\boldsymbol{k}} = \sum_{\sigma=-,+} \hat{\boldsymbol{c}}_{\boldsymbol{k}\sigma} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \ . \tag{3.14}$$

This corresponds to taking the real part of $2\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}$. Inspired by the total energy of the free electromagnetic field (3.9), we reformulate the Hamiltonian operator (3.13) in terms of the real part of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}$:

$$\begin{aligned} \hat{H} &= \frac{\varepsilon_0}{(2\pi)^3} \int \frac{1}{2} \Big(\omega_{\boldsymbol{k}}^2 (\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} - \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} - \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger) \\ &\quad + c^2 |\boldsymbol{k}|^2 (\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger \cdot \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger) \Big) \, \mathrm{d}\boldsymbol{k} \\ &= \frac{\varepsilon_0}{(2\pi)^3} \int \frac{1}{2} \left(| - i\omega_{\boldsymbol{k}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + i\omega_{\boldsymbol{k}} \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger |^2 + c^2 | i\boldsymbol{k} \times (\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger) |^2 \right) \mathrm{d}\boldsymbol{k} \ . \end{aligned}$$

For the first equality we have used that $\omega_{\boldsymbol{k}} = c|\boldsymbol{k}|$ and for the second we have used that $\boldsymbol{k} \cdot (\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^\dagger) = 0$. This orthogonality relation follows from the definitions (3.12) given previously. The Coloumb gauge condition $\nabla \cdot \boldsymbol{A} = 0$ requires a similar orthogonality relation $\boldsymbol{k} \cdot \boldsymbol{c}_{\boldsymbol{k}\sigma} = 0$ for all $\boldsymbol{k}$. This means that, so far, our trial solution (3.14) fulfils the requirements of the Coulomb gauge.

If we make the following replacement

$$\hat{a}_{\boldsymbol{k}\sigma} |\Psi\rangle \quad \mapsto \quad a_{\boldsymbol{k}\sigma}(t) = a_{\boldsymbol{k}\sigma}(0) e^{-i\omega_{\boldsymbol{k}\sigma} t} \ ,$$

the conjugate operators become the complex conjugates and $\boldsymbol{\alpha}_{\boldsymbol{k}} + \boldsymbol{\alpha}_{\boldsymbol{k}}^*$ becomes the Fourier transform of the vector potential in our trial solution. The shift to wave amplitudes is another simplification of the quantum field theory. It corresponds to using the mean number of photons all with the same energy and momentum when specifying the intensity of the electromagnetic field [Martin and Rothen 2004]. Inserting in the expression for the Hamiltonian, we get the total energy

$$E = \frac{\varepsilon_0}{2(2\pi)^3} \int \left(\left| \frac{\partial}{\partial t} (\boldsymbol{\alpha}_{\boldsymbol{k}} + \boldsymbol{\alpha}_{\boldsymbol{k}}^*) \right|^2 + c^2 |\nabla \times (\boldsymbol{\alpha}_{\boldsymbol{k}} + \boldsymbol{\alpha}_{\boldsymbol{k}}^*)|^2 \right) \mathrm{d}\boldsymbol{k} \ . \tag{3.15}$$

Note that we do not get an operator $\hat{H}$ as the result, but rather the total energy E, because we replaced the quantum operators by classical waves. Since $\boldsymbol{c_k} = \boldsymbol{\alpha_k} + \boldsymbol{\alpha_k^*}$ is real, we can use the property (3.11) of the integral over the squared vector potential and get $E = E_M$ (compare Equations 3.15 and 3.9). Our trial solution was indeed a success. Consequently we can quantize the electromagnetic field by using an operator version of the vector potential $\boldsymbol{A}$ defined by

$$\hat{\boldsymbol{A}}(\boldsymbol{x}, t) = \frac{1}{(2\pi)^3} \int \left(\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}(t) e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} + \hat{\boldsymbol{\alpha}}_{\boldsymbol{k}}^{\dagger}(t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}} \right) \mathrm{d}\boldsymbol{k} \ . \tag{3.16}$$

The vector potential governs the behaviour of the free electromagnetic field, while the operator version $\hat{\boldsymbol{A}}$ governs the behaviour of a free photon field since it describes the Hamiltonian operator. Then the conclusion for the free field is that when the number of photons is plenty and when their energy is evenly distributed, Maxwell's equations agree with the mean effects of the quantum theoretical approach. We also conclude that decomposition of field vectors into two polarisation components entirely agrees with the quantum theory.

3.2 The Free Charge Field

Quantum electrodynamics concern interaction of the electromagnetic field with electrons and positrons [Dirac 1966]. We have just specified the behaviour of the electromagnetic field in free space. When materials are introduced things get more complicated. Then the Hamiltonian operator needs to be expanded by the Hamiltonian of the charge field and the interaction Hamiltonian such that

$$\hat{H} = \hat{H}_M + \hat{H}_D + \hat{H}_I \ ,$$

where the subscript M is for Maxwell, D is for Dirac, and I is for interaction. If we specify operator versions of the electric and magnetic field vectors using the operator version of the vector potential (3.16), we have

$$\hat{H}_M = \frac{\varepsilon_0}{2} \int (|\hat{\boldsymbol{E}}|^2 + c^2 |\hat{\boldsymbol{B}}|^2) \, \mathrm{d}\boldsymbol{x} \ . \tag{3.17}$$

Unfortunately we cannot use the same creation and annihilation operators for all types of quantum particles. If a particle has integral spin, it is called a *boson*, and for this type of particle the operators $\hat{a}_n^{\dagger}$ and $\hat{a}_n$ are valid. If the spin is a half-integer, the particle is called a *fermion*, and in this case we need two sets of creation and annihilation operators: One set for the particle itself, $\hat{b}_n^{\dagger}$ and $\hat{b}_n$, and one set for its antiparticle, $\hat{d}_n^{\dagger}$ and $\hat{d}_n$. Being spin one, photons

are bosons. Electrons, protons, and neutrons are all spin one-half and are, therefore, fermions. The fermion creation and annihilation operators follow the *Pauli exclusion principle* which states that only one electron is allowed in each energy state (that is, $\hat{b}_n^\dagger \hat{b}_n$ and $\hat{d}_n^\dagger \hat{d}_n$ are both either 0 or 1).

The energy of a fermion is given by the relativistic formula $E^2 = p^2 + m^2$, where $p = |\boldsymbol{p}|$ is the momentum and m is the rest mass of the particle. This means that there is both a positive and a negative energy solution $E = \pm\sqrt{p^2 + m^2}$ for the harmonic oscillator (3.5) which is supposed to describe the particle. To make this conception work, we will have to think of the physical vacuum as a state in which all negative-energy electron states are "filled". In other words the charge field gives rise to negative zero-point energy whereas the electromagnetic field gives rise to positive zero-point energy. The Hamiltonian operator for the Dirac field is [Milonni 1994]

$$\hat{H}_D = \sum_n E_n(\hat{b}_n^\dagger \hat{b}_n + \hat{d}_n^\dagger \hat{d}_n - 1) \ .$$

Because we have both positive and negative angular momenta ($-\hbar/2$ and $+\hbar/2$) as well as positive- and negative-energy solutions, we need four component vectors (so called *spinors*) to capture the wave equation of a free electron. Suppose we denote the positive-energy solutions $\psi_{n+}(\boldsymbol{x})$. These (spinors) will have the third and fourth components equal to zero. The negative-energy solutions $\psi_{n-}(\boldsymbol{x})$ will have the first and second components equal to zero. This gives the following wave expansion:

$$\psi(\boldsymbol{x}, t) = \sum_n (b_n(t)\psi_{n+}(\boldsymbol{x}) + d_n^* \psi_{n-}(\boldsymbol{x})) \ , \tag{3.18}$$

where $b_n(t) = b_n(0)e^{-iE_n t}$ and $d_n(t) = d_n(0)e^{-iE_n t}$.

When we replace b_n and d_n in Equation 3.18 by the quantization operators $\hat{b}_n$ and $\hat{d}_n$, it is clear that the probability amplitude function becomes a new operator $\hat{\psi}(\boldsymbol{x}, t)$. Again we refer to $\hat{\psi}^\dagger$ and $\hat{\psi}$ as the creation and annihilation operators, respectively, but their meaning is different from that of $\hat{b}_n^\dagger$, $\hat{d}_n^\dagger$, $\hat{b}_n$, and $\hat{d}_n$ in the sense that they only include one position in the particle configuration spaces that are added to or removed from the system. You could say that they create or annihilate particles at the position $\boldsymbol{x}$, but the truth is that the particle is only fully created once the operator has been integrated over all possible particle positions. In other words,

$$\begin{aligned} \hat{b}_n(t) &= \int \hat{\psi}(\boldsymbol{x}, t)\psi_{n+}^*(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ \hat{d}_n(t) &= \int \hat{\psi}(\boldsymbol{x}, t)\psi_{n-}^*(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \ , \end{aligned}$$

which follows from Equation 3.18 because the amplitudes ψ_n for the different energy states are orthonormal:

$$\int \psi_m(\boldsymbol{x})\psi_n^*(\boldsymbol{x'})\,\mathrm{d}\boldsymbol{x} = \delta_{mn} \ .$$

Here m and n include both positive- and negative-energy indices and δ_{mn} is the Kronecker delta which is 1 for $m = n$ and 0 otherwise.

To handle the four-component spinors resulting from the creation and annihilation operators $\hat{\psi}^\dagger$ and $\hat{\psi}$, we need a set of 4×4 matrices which apply the necessary coefficients to the different components. For this purpose we use the Dirac representation of the *Pauli spin matrices* [Dirac 1930]:

$$\chi_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad , \quad \chi_2 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & i & 0 \\ 0 & -i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix}$$

$$\chi_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad , \quad \beta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \ .$$

These matrices give us the opportunity to write the Hamiltonian operator for the Dirac field as follows [Milonni 1994]:

$$\hat{H}_D = \int \hat{\psi}^\dagger(-i\boldsymbol{\chi} \cdot \nabla + \beta m)\hat{\psi}\,\mathrm{d}\boldsymbol{x} \ .$$

Note that the vector $\boldsymbol{\chi} = (\chi_1, \chi_2, \chi_3)$ of three 4×4 matrices is needed because the gradient of a spinor will result in a vector of three four-component spinors. If we insert this Hamiltonian operator in the Schrödinger equation (3.4), we have the equation governing the second of the three basic actions ("an electron goes from place to place").

3.3 Interaction of the Fields

The definitions of the creation and annihilation operators means that the combined operator $\hat{\psi}^\dagger\hat{\psi}$ does not count the total number of electrons versus the number of positrons. Instead it counts the number of electrons versus the number of positrons at a specific position in space. This corresponds to the charge density ρ. The quantized version of the charge density is then

$$\hat{\rho}(\boldsymbol{x}, t) = -q_e\,\hat{\psi}^\dagger(\boldsymbol{x}, t)\hat{\psi}(\boldsymbol{x}, t) \ ,$$

where $-q_e = -1.602 \cdot 10^{-19}\,\mathrm{C}$ is the charge on a single electron. Using the Pauli spin matrices, the $\hat{\psi}$ operators also give us a simple way of quantizing the current density $\boldsymbol{j}$ [Milonni 1994]:

$$\hat{\boldsymbol{j}}(\boldsymbol{x},t) = q_e\,\hat{\psi}^\dagger(\boldsymbol{x},t)\,\boldsymbol{\chi}\,\hat{\psi}(\boldsymbol{x},t)\ . \tag{3.19}$$

To ensure charge conservation, we have the condition

$$\frac{\partial\hat{\rho}}{\partial t} + \nabla\cdot\hat{\boldsymbol{j}} = 0\ . \tag{3.20}$$

In the Coulomb gauge, where $\nabla\cdot\boldsymbol{A} = 0$, another way to write Maxwell's equations is [Feynman et al. 1964, Sec. 18-6, but using the Coulomb gauge instead of the Lorentz gauge]

$$\nabla^2\phi = -\rho/\varepsilon_0 \tag{3.21}$$

$$c^2\nabla^2\boldsymbol{A} - \frac{\partial^2\boldsymbol{A}}{\partial t^2} = -\frac{\boldsymbol{j}}{\varepsilon_0} + \nabla\frac{\partial\phi}{\partial t}\ . \tag{3.22}$$

This reveals how the creation and annihilation operators $\hat{\psi}^\dagger$ and $\hat{\psi}$ link the charge field to the electromagnetic field. Consider the right-hand side of the second equation (3.22):

$$\boldsymbol{j}' = -\frac{\boldsymbol{j}}{\varepsilon_0} + \nabla\frac{\partial\phi}{\partial t}\ .$$

Taking the divergence on both sides, applying the equation (3.21) which involves the scalar potential, and using the charge conservation condition (3.20), we see that

$$\nabla\cdot\boldsymbol{j}' = -\frac{1}{\varepsilon_0}\left(\nabla\cdot\boldsymbol{j} + \frac{\partial\rho}{\partial t}\right) = 0\ .$$

Meaning that $\boldsymbol{j}' = \boldsymbol{j}_\perp$ is the transverse component of the current density.

Fourier expansion of the transverse component of the current density $\boldsymbol{j}_\perp$, gives the Fourier coefficients

$$\boldsymbol{j}^\perp_{\boldsymbol{k}}(t) = \int \boldsymbol{j}_\perp(\boldsymbol{x},t)e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\,\mathrm{d}\boldsymbol{x}\ .$$

Then by applying the Fourier transformation to Equation 3.22, we obtain

$$c^2|\boldsymbol{k}|^2\boldsymbol{c}_{\boldsymbol{k}} + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{c}_{\boldsymbol{k}} = \frac{\boldsymbol{j}^\perp_{\boldsymbol{k}}}{\varepsilon_0}\ , \tag{3.23}$$

where $\boldsymbol{c}_{\boldsymbol{k}}$ are the Fourier coefficients of the vector potential (cf. Equation 3.10). This result corresponds to the equation of motion for a forced oscillator. Regarding each energy state (or wave mode given by $\boldsymbol{k}$ and the polarisation σ)

as a forced oscillator, the energy of the classical electromagnetic field (in the presence of matter) is [Martin and Rothen 2004]

$$\begin{aligned} E_{MI} &= E_M - \frac{1}{(2\pi)^3}\int (\boldsymbol{\alpha}_{\boldsymbol{k}}^*(0)\cdot \boldsymbol{j}_{\boldsymbol{k}}(t) + \boldsymbol{\alpha}_{\boldsymbol{k}}(0)\cdot \boldsymbol{j}_{\boldsymbol{k}}^*(t))\,\mathrm{d}\boldsymbol{k} \\ &= E_M - \int \boldsymbol{j}(\boldsymbol{x},t)\cdot \boldsymbol{A}(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x}\ . \end{aligned} \tag{3.24}$$

If we replace current density $\boldsymbol{j}$ and vector potential $\boldsymbol{A}$ with their operator counterparts (insert hats) in this equation (3.24), we get a complete analogy between the classical field and the quantum field. It makes perfect sense that the quantized current density $\hat{\boldsymbol{j}}$ governs the creation and annihilation of photons over time. Letting $\hat{\boldsymbol{A}}$ denote the vector potential operator (in the Coulomb gauge) at time $t = 0$ and using Equation 3.19, we have

$$\hat{H}_I = -\int \hat{\boldsymbol{j}}(\boldsymbol{x},t)\cdot \hat{\boldsymbol{A}}(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = -\int q_e\,\hat{\psi}^\dagger \boldsymbol{\chi}\cdot \hat{\boldsymbol{A}}\hat{\psi}\,\mathrm{d}\boldsymbol{x}\ .$$

This is the Hamiltonian operator governing the third of the three basic actions ("an electron emits or absorbs a photon")

In total a system involving photons, electrons, positrons, and their interactions is described by the Hamiltonian operator

$$\begin{aligned} \hat{H} &= \hat{H}_D + \hat{H}_M + \hat{H}_I \\ &= \int \left(\hat{\psi}^\dagger(-i\boldsymbol{\chi}\cdot\nabla + \beta m)\hat{\psi} + \frac{\varepsilon_0}{2}(|\hat{\boldsymbol{E}}|^2 + c^2|\hat{\boldsymbol{B}}|^2) - q_e\,\hat{\psi}^\dagger\boldsymbol{\chi}\cdot\hat{\boldsymbol{A}}\hat{\psi}\right)\mathrm{d}\boldsymbol{x}\ . \end{aligned} \tag{3.25}$$

If we want to include more effects, such as the binding of electrons by protons, the Hamiltonian operator gets more complicated (in the case of an external binding potential we would have to add a term $\hat{\psi}^\dagger V\hat{\psi}$ under the integration). The Hamiltonian is good for describing the relations between quantum and classical electrodynamics, but it is not very practical for deriving the behaviour of quantum particle systems. Based on perturbation theory Feynman [1949a; 1949b] found a much more convenient formulation of quantum electrodynamics. In Feynman's approach there is no need for quantization or specification of the Hamiltonian, in his own words [Feynman 1949a, p. 749]:

> The main principle is to deal with the solutions to the Hamiltonian differential equations rather than with the equations themselves.
>
> [...]
>
> [...] we imagine the entire space-time history laid out, and that we just become aware of it successively. In a scattering problem this over-all view of the complete scattering process is similar to the S-matrix viewpoint of Heisenberg. The temporal order of events during the scattering, which is analyzed in such detail by the Hamiltonian differential equation, is irrelevant.

The S matrix describes the amplitude $\langle \Psi_2 | S | \Psi_1 \rangle$ that a system state Ψ_1 in the infinite past turns into the system state Ψ_2 in the infinite future. If we describe the S matrix by an appropriate expansion of integrals over space and time, we become aware of another part of space-time history for every term that we include in our evaluation of the expansion. This is Feynman's way of handling the infinite number of degrees of freedom in the system. Unfortunately this approach is more difficult to relate to the classical electrodynamics, therefore the standard formalism has been used in this chapter.

3.4 A Quantum Field Simulator

In summary the simplifications imposed on a system when we go from the quantum field to the classical electromagnetic field are the following:

- Field energy and charge is distributed evenly across the frequency spectrum.
- Emission and absorption occurs continuously not in quanta.
- The amplitudes of the classical field vectors ($\boldsymbol{A}$, $\boldsymbol{E}$, $\boldsymbol{B}$, $\boldsymbol{j}$) replaces the properties of the individual particles by mean values.
- Phenomena related to zero-point energy are neglected.

There are many ways of modifying the classical electromagnetic field to help some of these issues. We could use a discrete set of wave modes, use different components for the polarisations of light and matter, and add some zero-point energy. But if we zoom in close enough, the classical theory will always have difficulties. The fundamental difference between the two approaches is that in the classical theory we get the fraction of energy ending up in one place or another, while in the quantum theory we get probabilities telling us how often photons will end up in these places.

In the context of light transport simulation for graphics, it is difficult to think of a case where the limitations of using Maxwell's equations would be of any significance. Most of the time we consider scenarios where photons are plenty, and we do not need to worry about anything other than the mean effects. Nevertheless, we do use the blackbody emission spectrum, and if we want to determine the optical properties of materials based on their microscopic composition, the quantum theories easily become important. We will return to the microscopic composition of materials in Part II. With respect to light propagation it could

be interesting to construct a "particle tracer" based on probabilities rather than energy fractions. How would one go about constructing such a tracer? Would it be possible to construct a rendering algorithm based on quantum electrodynamics? Let us try to formulate what I refer to as a *quantum field simulator*.

Before we start constructing a rendering algorithm, we first need to know what the output ought to be. What we want is an image on a computer screen. In other words, we want the screen to emit the light that would have been reflected or transmitted from an object in the real world towards our eyes. The light emitted by a computer screen is described by a set of colour values with a red, a green, and a blue component (RGB). The computer is imitating the trichromatic colour vision of the eyes when making its output. As in the Young-Helmholtz theory the RGB colour values are related to the wavelengths of the light field that we want the screen to emit. And that is again the light field that we want the eyes to receive. According to the quantum theory of light, the wavelengths in a beam of light are determined by the energies of the photons in the field. The intensity of the light, that is, the magnitude of the colour values, is determined by the number of photons arriving per second.

A rendering algorithm works as follows: We have a mathematical model of the scene (or system) that we want to render. Starting from the eyes or the light sources, we seek or trace the light in the scene. The image is a planar surface placed somewhere close to the eye (normal to the viewing direction). RGB colour values are computed for each small part (pixel) of the image plane using the light that reaches that part.

To make a rendering algorithm based on quantum electrodynamics, we would first have to find all the relevant paths that light can take from the source to the eye. The scene should be divided into surface patches or voxels (volume elements), and the paths specified by a sequence of patches or voxels that the light travels across from surface to eye. Then we find the probability amplitudes for photons to take each of the alternative paths. Wherever a photon path meets a patch or voxel of material, we should include the possibility of emission or absorption of a photon by an electron. Using the probability amplitudes for the different paths that a photon can take, the probability that a photon of angular momentum (polarisation) σ and energy E_n reaches some part of the image plane is found according to the rules given at the very beginning of this chapter (p. 45). Considering the number of photons emitted per second and the probability that a photon of a particular polarisation and energy will hit a specific part of the image, we get the intensity and frequency spectrum arriving at every pixel for every second. How do we find the probability amplitudes? We solve the Schrödinger equation (3.4) using the Hamiltonian operator (3.25) or, alternatively, we use Feynman's approach.

This sketch of a scheme for implementing a quantum field simulator, reveals that we would have to make many shortcuts to make it work in graphics. Modelling all the electrons of all the materials appearing in a scene is an insurmountable task. Some sort of macroscopic representation of materials is necessary. Nevertheless, the concept of using probabilities for every path through a scene is perhaps not such a bad idea. It is possible at a more macroscopic level, and it would make it easier to describe phenomena such as diffraction. The implementation of a quantum field simulator is left as a curio for future work. For the time being, we will move towards a more macroscopic description of nature. To take the first step, we look at Maxwell's equations in the following chapter.

Chapter 4

Electromagnetic Radiation

Whether it be molecules, the waves of the sea or myriads of stars, elements of nature form overall structures.

Peter Haarby describing Inge Lise Westman's paintings

In the previous chapter we found a number of correspondences between quantum fields and Maxwell's equations. In particular, we found that the electromagnetic field vectors capture the mean effects of quantum particles. The Maxwell equations that we found correspondences to are sometimes referred to as the microscopic Maxwell equations. These equations only involve the electric and magnetic field vectors ($\boldsymbol{E}$ and $\boldsymbol{B}$), and the charge and current densities ($\boldsymbol{j}$ and ρ). The charge and current densities are material specific quantities referring to the behaviour of the electrons in the material. The field vectors are a more macroscopic way of describing light.

What we are really interested in is the progress of energy in a field. Therefore we introduce the Poynting vector in this chapter (Sec. 4.1). The Poynting vector is a quantity describing the energy flow in an electrodynamic field. Based on this quantity, we try to say something about the propagation of electromagnetic energy. This leads to a justification for the inverse square law of radiation and a formal solution for the microscopic equations. Afterwards we move to a more macroscopic description of charges and currents (Sec. 4.2). This is necessary since it is difficult to model every atom in a graphics scene. More macroscopic material properties are introduced, and, with those, we obtain the so-called macroscopic Maxwell equations. The next step is to investigate the dif-

ferent wave functions that we can use as solutions for the macroscopic equations (Sec. 4.3). Using some simplifying assumptions, we arrive at plane waves as a simple solution. Finally, we describe some wave theory which is used extensively in graphics. In particular, we derive the law of reflection, the law of refraction, and the Fresnel equations for reflection and transmission (Sec. 4.4). As an additional feature we show that variants of these laws and formulae are also valid for the important case of inhomogeneous waves (almost any wave propagating in an absorbing material is inhomogeneous).

4.1 Microscopic Maxwell Equations

From the theory of quantum electrodynamics we understand that the interaction of photons and electrons to a certain extent agrees with Maxwell's equations:

$$c^2 \nabla \times \boldsymbol{B} = \frac{\boldsymbol{j}}{\varepsilon_0} + \frac{\partial \boldsymbol{E}}{\partial t} \tag{4.1}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{4.2}$$

$$\nabla \cdot \boldsymbol{E} = \rho/\varepsilon_0 \tag{4.3}$$

$$\nabla \cdot \boldsymbol{B} = 0 \ , \tag{4.4}$$

where $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic field vectors, $\boldsymbol{j}$ and ρ are the current and charge densities, c is the speed of light in vacuum, and ε_0 is the vacuum permittivity. The original version of Maxwell's equations included two additional vectors ($\boldsymbol{H}$ and $\boldsymbol{D}$) because Maxwell (and other early workers) were not aware of the internal structure of atoms. They did not know that charges are bound to atoms and that atomic magnetism is caused by circulating currents [Feynman et al. 1964, Sec. 32-2]. This means that they had to work with a more general set of equations than what is really needed to describe nature. Maxwell's old magnetic vector $\boldsymbol{H}$ proves to be useful in a more macroscopic description of the electromagnetic field. Therefore the field $\boldsymbol{H}$ is still used extensively and we will introduce it later in this chapter.

In our thought experiment of the previous chapter concerning a quantum field simulator for rendering realistic images, we reasoned that the intensity of RGB colour values is determined by the number of photons arriving per second. The relative amount of red, green, and blue is determined by the energies of the photons. When we reduce the quantum particles to waves and use only Maxwell's equations, we have to think of it a little differently. The number of photons arriving per second is rather the magnitude of the energy flux in the field and the energies of the photons are the wavelengths present in the field. When we

look at Maxwell's equations there is no quantity describing the energy of the field. We need such a quantity in order to relate the electromagnetic field to the RGB colour values of our image.

Simple arguments show that the loss of energy per unit time and per unit volume due to work done by the electromagnetic field is the quantity $\boldsymbol{E} \cdot \boldsymbol{j}$ [Feynman et al. 1964]. This was also known to Maxwell [1873, Vol. II, Chapter VI] and using his equation (4.1) involving the curl of the magnetic field, one can also write this quantity as

$$\boldsymbol{E} \cdot \boldsymbol{j} = \varepsilon_0 c^2 \boldsymbol{E} \cdot (\nabla \times \boldsymbol{B}) - \varepsilon_0 \boldsymbol{E} \cdot \frac{\partial \boldsymbol{E}}{\partial t} \ .$$

Poynting [1884] essentially showed that another way to write it is

$$\boldsymbol{E} \cdot \boldsymbol{j} = -\nabla \cdot (\varepsilon_0 c^2 \boldsymbol{E} \times \boldsymbol{B}) - \frac{\partial}{\partial t} \left(\frac{\varepsilon_0}{2} \boldsymbol{E} \cdot \boldsymbol{E} + \frac{\varepsilon_0 c^2}{2} \boldsymbol{B} \cdot \boldsymbol{B} \right) \ .$$

Looking back at the equation (3.20) of charge conservation, this equation is remarkably similar. Only there is a loss of electromagnetic energy ($-\boldsymbol{E} \cdot \boldsymbol{j}$) whereas there is no loss of charge in the field. Considering this analogy, one defines the energy flux $\boldsymbol{S}$ of the field, also called Poynting's vector, and the energy density u of the field as follows:

$$\boldsymbol{S} = \varepsilon_0 c^2 \boldsymbol{E} \times \boldsymbol{B} \tag{4.5}$$

$$u = \frac{\varepsilon_0}{2} (|\boldsymbol{E}|^2 + c^2 |\boldsymbol{B}|^2) \ , \tag{4.6}$$

such that

$$\nabla \cdot \boldsymbol{S} + \frac{\partial u}{\partial t} = -\boldsymbol{E} \cdot \boldsymbol{j} \ . \tag{4.7}$$

In the previous chapter we saw how the expression for the energy density (4.6) agrees well with the quantized description of the energy in a photon field (cf. Equations 3.7 and 3.17). Thus we can use the intensity of Poynting's vector $|\boldsymbol{S}|$ to represent the intensity of the colour values in our renderer.

Knowing how to find the energy flux of the field, the next thing we need to know, is how to follow the propagation of the waves through a scene. How are waves of light emitted, how are they absorbed, how do they interact with matter? To start with emission, the formula for electromagnetic radiation by one individual point charge in free space is as follows [Feynman et al. 1963, Sec. 28-1]:

$$\boldsymbol{E} = \frac{q}{4\pi\varepsilon_0} \left(\frac{\vec{e}_{r'}}{r'^2} + \frac{r'}{c} \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\vec{e}_{r'}}{r'^2} \right) + \frac{1}{c^2} \frac{\mathrm{d}^2}{\mathrm{d}t^2} \vec{e}_{r'} \right) \tag{4.8}$$

$$\boldsymbol{B} = \vec{e}_{r'} \times \boldsymbol{E}/c \ , \tag{4.9}$$

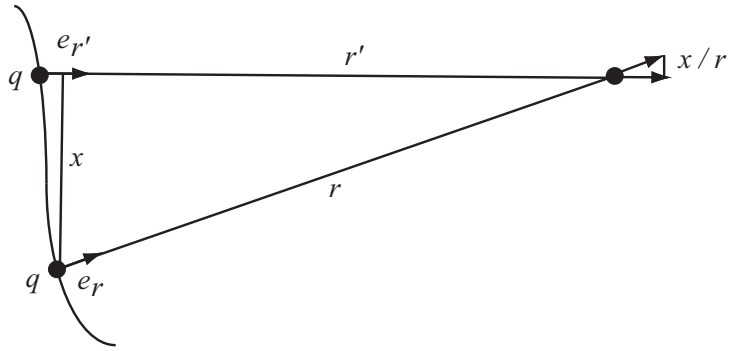


Figure 4.1: *A moving point charge q at a large distance $r \approx r'$ from an observer. This figure illustrates why the displacement of the vector $\vec{e}_{r'}(t)$ on the unit sphere is approximately equal to the displacement $x(t)/r$.*

where q is the charge and $\boldsymbol{r}(t) = r(t)\vec{e}_r(t)$ is the vector at time t from the charge toward the position we are considering in the field. The distance is denoted $r(t)$ and the unit vector describing the direction of $\boldsymbol{r}$ is denoted $\vec{e}_r(t)$. The distance and direction appearing in the formulae (4.8–4.9) are retarded such that $r' = r(t - r'/c)$ and $\vec{e}_{r'} = \vec{e}_r(t - r'/c)$. Note that these expressions are recursive. We cannot determine the electromagnetic field at an instance in time without knowing where the charges were at some time in the past.

If we look closer at the equation for radiation from a single charge (4.8), it reveals that the first two terms will vanish quickly as the distance to the charge increases. So let us assume that r is large and concentrate on the third term only:

$$\boldsymbol{E} = \frac{q}{4\pi\varepsilon_0 c^2}\frac{\mathrm{d}^2\vec{e}_{r'}}{\mathrm{d}t^2} \quad .$$

Let us assume that the charge moves slowly compared to the speed of light. Then the charge will move only a short distance from $\boldsymbol{r}'$ to $\boldsymbol{r}$. If we let x denote the length of $\boldsymbol{r} - \boldsymbol{r}'$ projected on a plane normal to $\boldsymbol{r}'$, the sine of the angle between $\boldsymbol{r}$ and $\boldsymbol{r}'$, that is, x/r, will approximately be the change in $\vec{e}_{r'}$. This is illustrated in Figure 4.1. Since the distance $r \approx r'$ is large, it is almost constant. Therefore the acceleration of $\vec{e}_{r'}$ is approximately $a'_\perp/r'$, where $a'_\perp$ is the retarded, projected acceleration of the charge itself. The prime denotes that it is retarded, which means that it is at the time $t - r'/c$, and the symbol $\perp$ denotes that it is projected on a plane normal to the viewing direction $\boldsymbol{r}'$. With this approximation, we have an expression for the magnitude of the electric field vector

$$|\boldsymbol{E}| = \frac{q}{4\pi\varepsilon_0 c^2 r'}a'_\perp \quad .$$

From the relation (4.9) between $\boldsymbol{E}$ and $\boldsymbol{B}$ we find the following (which is valid for energy radiated through a vacuum)

$$|\boldsymbol{B}| = |\boldsymbol{E}|/c \ , \tag{4.10}$$

Inserting in the expression for Poynting's vector (4.5), we find the that the intensity of the energy flux is

$$|\boldsymbol{S}| = \varepsilon_0 c |\boldsymbol{E}|^2 \ . \tag{4.11}$$

If we assume that the charge is accelerating in a direction forming the angle θ with $\boldsymbol{r}'$, we get the following intensity for a single charge far away from the point of observation:

$$|\boldsymbol{S}| = \frac{q^2}{16\pi^2 \varepsilon_0 c^3 r'^2} a'^2 \sin^2\theta \ , \tag{4.12}$$

where a' is the retarded acceleration of the charge. Note that the intensity of the radiation falls off with the square of the distance r' to the source. This is the justification for the inverse square law of radiation and now we are aware of the simplifying assumptions involved in its derivation. If we get close enough or if the charge is moving fast enough, the inverse square law is no longer valid. In that case we should use the general result given by Equations 4.11 and 4.8 (except for $r' = 0$).

The equation for a single point charge is interesting, but we need more than one charge to describe the scenes that we want to render in graphics. If we return to the vector potential $\boldsymbol{A}$ and scalar potential ϕ defined indirectly in the previous chapter (Equation 3.8), and choose the Lorentz gauge, we have

$$\boldsymbol{E} = -\nabla\phi - \frac{\partial \boldsymbol{A}}{\partial t} \quad , \quad \boldsymbol{B} = \nabla \times \boldsymbol{A} \quad , \quad \nabla \cdot \boldsymbol{A} = -\frac{1}{c^2}\frac{\partial \phi}{\partial t} \ . \tag{4.13}$$

Then Mawell's equations (4.1–4.4) become

$$\nabla^2 \boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\frac{\boldsymbol{j}}{\varepsilon_0 c^2} \tag{4.14}$$

$$\nabla^2 \phi - \frac{1}{c^2}\frac{\partial^2 \phi}{\partial t^2} = -\frac{\rho}{\varepsilon_0} \ . \tag{4.15}$$

These two equations are actually four differential equations of identical structure ($\boldsymbol{A}$ has three components). Formally they have the same solution. This general solution for Maxwell's equations at the time t is [Feynman et al. 1964, Sec. 21-3]

$$\boldsymbol{A}(\boldsymbol{x}, t) = \int \frac{\boldsymbol{j}(\boldsymbol{y}, t - r_{\boldsymbol{xy}}/c)}{4\pi\varepsilon_0 c^2 r_{\boldsymbol{xy}}} \, \mathrm{d}\boldsymbol{y} \tag{4.16}$$

$$\phi(\boldsymbol{x}, t) = \int \frac{\rho(\boldsymbol{y}, t - r_{\boldsymbol{xy}}/c)}{4\pi\varepsilon_0 r_{\boldsymbol{xy}}} \, \mathrm{d}\boldsymbol{y} \ , \tag{4.17}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are positions in space. The field vectors are obtained by insertion in the equations (4.13) which define the potentials. It is possible to derive the field of a single point charge (4.8–4.9) using this solution [Feynman et al. 1964, Chapter 21].

4.2 Macroscopic Maxwell Equations

Since we do not want to simulate the interaction of every charge with the field, we need some more macroscopic measures. Considering the general solution (4.16–4.17), macroscopic expressions for the charge and current densities (ρ and $\boldsymbol{j}$) seem to be the right way to go. Suppose we want to model a material with N atoms per unit volume. We model each atom as having just one general dipole moment $q\boldsymbol{d}$, where q is the magnitude of the charges in the atom and $\boldsymbol{d}$ is a vector denoting their separation. A dipole is two charges separated by a very short distance, but under a few assumptions any assembly of point charges approximately has a dipole potential. The assumptions are that the charges should be (a) located in a small limited region, (b) neutral as a whole, and (c) observed at a large distance [Feynman et al. 1964, Sec. 6-5]. The dipole moment per unit volume is called the *polarisation vector* and is given by

$$\boldsymbol{P} = Nq\boldsymbol{d} \ .$$

In this dipole approximation the polarisation vector is proportional to the electric field vector $\boldsymbol{E}$, we write

$$\boldsymbol{P} = \varepsilon_0 \chi_e \boldsymbol{E} \ ,$$

and refer to χ_e as the *electric susceptibility*. The charge and current densities due to the polarisation of a material are [Feynman et al. 1964, Sec. 10-3 and Sec. 32-2]

$$\rho_{\text{pol}} = -\nabla \cdot \boldsymbol{P} \quad , \quad \boldsymbol{j}_{\text{pol}} = \frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}t} \ . \tag{4.18}$$

If there are no charges or currents in free space and no magnetisation currents in the material, these charge and current densities are the only ones present. The polarisation vector is therefore a more macroscopic or phenomenological way of describing the charges and currents in a dielectric.

Magnetisation is not related to the charge density, "the magnetisation of materials comes from circulating currents within the atoms" [Feynman et al. 1964, Sec. 36-1]. To describe this, one introduces another macroscopic quantity: The *magnetisation vector* $\boldsymbol{M}$. It is defined indirectly by

$$\boldsymbol{j}_{\text{mag}} = \nabla \times \boldsymbol{M} \ . \tag{4.19}$$

Summing up the different terms in the charge and current densities, we have

$$\begin{aligned} \rho &= \rho_{\text{free}} + \rho_{\text{pol}} \\ \boldsymbol{j} &= \boldsymbol{j}_{\text{free}} + \boldsymbol{j}_{\text{pol}} + \boldsymbol{j}_{\text{mag}} \ . \end{aligned}$$

Inserting these in Maxwell's equations and moving the terms around, we get the following result:

$$\nabla \times (\varepsilon_0 c^2 \boldsymbol{B} - \boldsymbol{M}) = \boldsymbol{j}_{\text{free}} + \frac{\partial}{\partial t}(\varepsilon_0 \boldsymbol{E} + \boldsymbol{P}) \tag{4.20}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{4.21}$$

$$\nabla \cdot (\varepsilon_0 \boldsymbol{E} + \boldsymbol{P}) = \rho_{\text{free}} \tag{4.22}$$

$$\nabla \cdot \boldsymbol{B} = 0 \ . \tag{4.23}$$

If we introduce two additional field vectors:

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} \tag{4.24}$$

$$\boldsymbol{H} = \varepsilon_0 c^2 \boldsymbol{B} - \boldsymbol{M} \ , \tag{4.25}$$

we get the original Maxwell equations. Sometimes this version of the equations is referred to as the macroscopic Maxwell equations because they involve the phenomenological polarisation and magnetisation vectors ($\boldsymbol{P}$ and $\boldsymbol{M}$).

The magnetisation vector is often assumed to be proportional to $\boldsymbol{H}$ such that

$$\boldsymbol{M} = \chi_m \boldsymbol{H} \ , \tag{4.26}$$

where χ_m is the *magnetic susceptibility*. Then a rearrangement of Equation 4.25 gives

$$\boldsymbol{H} = \frac{\epsilon_0 c^2}{1 + \chi_m} \boldsymbol{B} \ .$$

This assumption is, however, only valid for a very limited set of magnetic materials. Another assumption we can make is that free charges only appear as conduction in a material. It has been found experimentally that metals produce a current with a density $\boldsymbol{j}$ proportional to $\boldsymbol{E}$ [Feynman et al. 1964, Sec. 32-6]. To model this relationship, we introduce another phenomenological quantity called the *conductivity* σ, such that

$$\boldsymbol{j}_{\text{free}} = \sigma \boldsymbol{E} \ .$$

To summarise these macroscopic or phenomenological material properties, we have

$$\boldsymbol{D} = \varepsilon_0 (1 + \chi_e) \boldsymbol{E} = \varepsilon \boldsymbol{E} \tag{4.27}$$

$$\boldsymbol{B} = (1 + \chi_m)/(\varepsilon_0 c^2) \boldsymbol{H} = \mu \boldsymbol{H} \tag{4.28}$$

$$\boldsymbol{j}_{\text{free}} = \sigma \boldsymbol{E} \ . \tag{4.29}$$

To shorten notation, these equations also introduce the permittivity ε and the permeability μ. Note that all these macroscopic material properties are independent of the direction of the field. For this reason they are often referred to as the *isotropic* material equations. When using these material equations, we cannot model as general a case as if we use Equations 4.20–4.23. If we use the indirect definitions of $\boldsymbol{P}$ and $\boldsymbol{M}$ (Equations 4.18 and 4.19), we do not have to make any simplifying assumptions about the material. Unfortunately the polarisation and magnetisation vectors are not easy to model, so in the following we will use the isotropic material equations.

Let us briefly get an idea about how the isotropic material properties relate to real-world materials [Born and Wolf 1999]: If σ is not negligibly small, the material is a *conductor* (which roughly means that it has some electrons that are not bound to any particular atom such that they are able to produce a "free" current). As an example metals are good conductors. If the material is not a conductor, it is called a *dielectric*. The electric properties are then determined solely by the permittivity ε. If μ differs appreciably from unity, the material is *magnetic*. In particular, if $\mu > 1$, the material is *paramagnetic*, while if $\mu < 1$ it is *diamagnetic*. The material properties are all wavelength dependent.

4.3 Time-Harmonic Solution and Plane Waves

If we look at the differential equations (4.14–4.15) which give rise to the general solution for the electromagnetic field, they reveal that the vector and scalar potentials have wave solutions at locations in space where there is no charge or current density (where ρ and $\boldsymbol{j}$ are zero). Let us represent the solution as time-harmonic plane waves. Then the potentials have the form

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{x},t) &= \mathrm{Re}\left(\boldsymbol{A}_0 e^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})}\right) \\ \phi(\boldsymbol{x},t) &= \mathrm{Re}\left(\phi_0 e^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})}\right) \quad , \end{aligned}$$

where $\boldsymbol{k}$ is the *wave vector* and Re takes the real part of a complex quantity. If we insert these solutions in the expressions (4.13) for $\boldsymbol{E}$ and $\boldsymbol{B}$, we again get time-harmonic plain waves. This plane wave solution is valid for radiation in free space, that is, when ρ and $\boldsymbol{j}$ are zero. Let us try to figure out if it is also valid in a more general case.

Having stepped away from the quantum theories, we can safely express the field vectors in terms of Fourier transforms (the assumption is that the waves have a

continuous range of frequencies):

$$\boldsymbol{F}(\boldsymbol{x},t) = \frac{1}{\pi}\int_0^\infty \mathcal{F}(\boldsymbol{x},\omega)e^{-i\omega t}\,\mathrm{d}\omega \ .$$

This is the argument why we are allowed to represent the field vectors as a superposition of time-harmonic functions:

$$\boldsymbol{F}(\boldsymbol{x},t) = \mathrm{Re}\left(\boldsymbol{F}_0(\boldsymbol{x})e^{-i\omega t}\right) \ .$$

In this representation both $\boldsymbol{D}$ and $\boldsymbol{E}$ are functions of $e^{-i\omega t}$. From the material equation (4.27) we then conclude that the permittivity ε does not depend on time. Similarly the permeability μ does not depend on time.

The time-harmonic representation of the electromagnetic field is very convenient. Therefore let us write Maxwell's equations using complex time-harmonic vector functions. To make it clear that the field vectors are complex, we follow the notation of Bohren and Huffman [1983] and use the subscript c. As an example the time-harmonic representation of the electric field vector is

$$\boldsymbol{E}_c = \boldsymbol{E}_0(\boldsymbol{x})e^{-i\omega t} \ ,$$

where it is understood that we obtain the physical electric field vector by taking the real part $\boldsymbol{E} = \mathrm{Re}(\boldsymbol{E}_c)$. With a loss of generality that is of no significance in a graphics context, we neglect charges moving freely through empty space, that is, we set $\rho_{\mathrm{free}} = 0$. Using the isotropic material equations (4.27–4.29) and the time independence of ε and μ, we get the following time-harmonic version of the macroscopic Maxwell equations:

$$\begin{aligned}
\nabla \times \boldsymbol{H}_c &= (\sigma - i\omega\varepsilon)\boldsymbol{E}_c && (4.30)\\
\nabla \times \boldsymbol{E}_c &= i\omega\mu\boldsymbol{H}_c && (4.31)\\
\nabla \cdot (\varepsilon\boldsymbol{E}_c) &= 0 && (4.32)\\
\nabla \cdot (\mu\boldsymbol{H}_c) &= 0 \ . && (4.33)
\end{aligned}$$

Note that we have packed most of the important charges and currents into the $\boldsymbol{H}_c$ vector and the material properties. By insertion of the plane wave expressions

$$\boldsymbol{E}_c(\boldsymbol{x},t) = \boldsymbol{E}_0 e^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})} \quad , \quad \boldsymbol{H}_c(\boldsymbol{x},t) = \boldsymbol{H}_0 e^{-i(\omega t - \boldsymbol{k}\cdot\boldsymbol{x})} \ , \qquad (4.34)$$

we observe that plane waves do not in general satisfy the conditions. But they do satisfy them if we assume that the material properties are independent of position, that is, if the material is *homogeneous*. Thus plane waves are not only a solution in the free electromagnetic field, but also when we use the isotropic

material equations and assume that the material is homogeneous. Inserting the plane wave solution, we get the following Maxwell equations:

$$\begin{aligned} \boldsymbol{k} \times \boldsymbol{H}_0 &= -\omega(\varepsilon + i\sigma/\omega)\boldsymbol{E}_0 && (4.35)\\ \boldsymbol{k} \times \boldsymbol{E}_0 &= \omega\mu\boldsymbol{H}_0 && (4.36)\\ \boldsymbol{k} \cdot \boldsymbol{E}_0 &= 0 && (4.37)\\ \boldsymbol{k} \cdot \boldsymbol{H}_0 &= 0 \ . && (4.38) \end{aligned}$$

This reveals that Maxwell's equations require plane waves to satisfy the following conditions:

$$\boldsymbol{k} \cdot \boldsymbol{E}_0 = \boldsymbol{k} \cdot \boldsymbol{H}_0 = \boldsymbol{E}_0 \cdot \boldsymbol{H}_0 = 0 \quad (4.39)$$

$$\boldsymbol{k} \cdot \boldsymbol{k} = \omega^2\mu(\varepsilon + i\sigma/\omega) \ , \quad (4.40)$$

where all the vectors may be complex and $\varepsilon_c = \varepsilon + i\sigma/\omega$ is sometimes called the complex permittivity (or the complex dielectric constant). The latter equation is particularly interesting, it denotes the relation between material and wave propagation.

Let us take a look at the real and imaginary parts of the wave vector $\boldsymbol{k}$. We write

$$\boldsymbol{k} = \boldsymbol{k}' + i\boldsymbol{k}'' = k'\vec{s}' + ik''\vec{s}'' \ ,$$

where $k' = |\boldsymbol{k}'|$ and $k'' = |\boldsymbol{k}''|$ such that $\vec{s}'$ and $\vec{s}''$ are unit vectors in the direction of real and imaginary part of the wave vector respectively. If the real part of the wave vector $\boldsymbol{k}'$ is parallel to the imaginary part $\boldsymbol{k}''$, the *wave* is said to be *homogeneous*. Otherwise it is *inhomogeneous*. Of course, $\boldsymbol{k}'' = \boldsymbol{0}$ is parallel to any vector, why a wave is homogeneous if $\boldsymbol{k}$ is real-valued. If $\boldsymbol{k}$ is complex, the exponential term of the plane wave expressions (4.34) is as follows

$$e^{i\,\boldsymbol{k}\cdot\boldsymbol{x}} = e^{i\,\boldsymbol{k}'\cdot\boldsymbol{x}} e^{-\boldsymbol{k}''\cdot\boldsymbol{x}} \ .$$

Here we may observe that $\boldsymbol{k}'$ is the vector normal to the surface of constant phase and $\boldsymbol{k}''$ is normal to the surface of constant amplitude. The phase velocity is then $v = \omega/k'$ and the amplitude is damped (or decays) in the direction $\vec{s}''$ at the rate k''.

If we consider the relation (4.40) describing the rule for propagation of plane waves in homogeneous matter, it is obvious that a phenomenological quantity with the following definition is convenient:

$$n = n' + in'' = c\sqrt{\mu(\varepsilon + i\sigma/\omega)} \ . \quad (4.41)$$

It is called the (complex) index of refraction, or refractive index. If we insert it in Equation 4.40, we get

$$\boldsymbol{k} \cdot \boldsymbol{k} = \boldsymbol{k}' \cdot \boldsymbol{k}' - \boldsymbol{k}'' \cdot \boldsymbol{k}'' + i2\boldsymbol{k}' \cdot \boldsymbol{k}'' = \frac{\omega^2}{c^2} n^2 \ . \quad (4.42)$$

For materials that are not strong absorbers, $\boldsymbol{k}'' \cdot \boldsymbol{k}''$ will be so small that we can neglect it. Then if we equate the real parts (and assume that the index of refraction is positive), we get

$$k' \approx \frac{\omega}{c} n' \ .$$

Hence, the real part of the refractive index is nearly the ratio of the speed of light in vacuum to the phase velocity $n' \approx k'c/\omega = c/v$. Equating the imaginary parts, has the result

$$2k'k'' \cos\theta = \frac{\omega^2}{c^2} 2n'n'' \ ,$$

where θ is the angle between the real and imaginary parts of the wave vector ($\boldsymbol{k}'$ and $\boldsymbol{k}''$). Using the approximate expression for k', we have

$$k'' \approx \frac{\omega}{c} \frac{n''}{\cos\theta} = \frac{2\pi}{\lambda} \frac{n''}{\cos\theta} \ .$$

where λ is the wavelength *in vacuum*. This means that the imaginary part n'' is an expression related to the absorption of light in the material.

The index of refraction is a nice way to sum up all the material properties, and indeed it is a quantity which is measured as one of the key optical properties of materials. We will return to the optical properties of materials in Part II.

Considering the energy of the field, we know that it is the magnitude of the Poynting vector:

$$|\boldsymbol{S}| = |\varepsilon_0 c^2 \boldsymbol{E} \times \boldsymbol{B}| = \varepsilon_0 c^2 |\boldsymbol{E} \times \boldsymbol{B}| \ .$$

For the plane wave solution, we have

$$|\boldsymbol{S}| = \varepsilon_0 c^2 \mu |\mathrm{Re}(\boldsymbol{E}_c) \times \mathrm{Re}(\boldsymbol{H}_c)| \ ,$$

where we have used one of the isotropic material equations (4.28) and the plane wave expressions (4.34). When the plane wave expressions are inserted, we get

$$|\boldsymbol{S}| = \varepsilon_0 c^2 \mu \left| \mathrm{Re}(\boldsymbol{E}_0 e^{-i(\omega t - \boldsymbol{k}' \cdot \boldsymbol{x})}) \times \mathrm{Re}(\boldsymbol{H}_0 e^{-i(\omega t - \boldsymbol{k}' \cdot \boldsymbol{x})}) \right| e^{-2k'' \boldsymbol{s}'' \cdot \boldsymbol{x}} \ .$$

Since the exponential terms which involve ωt and $\boldsymbol{k}' \cdot \boldsymbol{x}$ are only oscillations, it follows that $2k''$ is the exponential attenuation of the energy flux as the wave propagates through the material. This attenuation is called the *absorption coefficient* and in graphics we use the symbol σ_a (which should not be confused with the conductivity σ). The relationship between the imaginary part of the refractive index and the absorption coefficient is

$$\sigma_a = 2k'' \approx \frac{4\pi n''}{\lambda \cos\theta} \ , \tag{4.43}$$

where $\cos\theta = 1$ for homogeneous plane waves. After explaining all these quantities, we have come from a description of radiation from point charges at a microscopic level to a description of absorption of plane waves at a macroscopic level. We have not yet discussed how the wave changes when it meets a surface. This is the subject of the following section.

4.4 Reflection and Refraction

Let us consider a plane wave incident on a smooth surface. Due to the photon spin discussed in Chapter 3, it is convenient to resolve all the waves we deal with into two independent plane wave components. The wave components we choose are the wave with the electric vector perpendicular to the plane of incidence, $\perp$-polarised light, and the wave with the electric vector parallel to the plane of incidence, $\parallel$-polarised light. From experience we know that light incident on a smooth surface gives rise to two waves: A reflected and a transmitted wave. In the following we denote the incident wave by the subscript i, the reflected by the subscript r, and the transmitted by the subscript t. The boundary conditions given by Maxwell's equations require that the tangential component of the electric vector is continuous across the boundary of the surface. The $\perp$-polarised component of the electric vector is clearly tangent to the surface at the point of incidence, therefore at the boundary:

$$\boldsymbol{E}_{\perp i} + \boldsymbol{E}_{\perp r} = \boldsymbol{E}_{\perp t} \ .$$

This must hold at all times and no matter where we place the point of incidence in space. Suppose we place the point of incidence at the origin of our coordinate system (where $\boldsymbol{x} = \boldsymbol{0}$), then

$$\boldsymbol{E}_{0i}^{\perp} e^{-i\omega_i t} + \boldsymbol{E}_{0r}^{\perp} e^{-i\omega_r t} = \boldsymbol{E}_{0t}^{\perp} e^{-i\omega_t t} \ .$$

This is true only if

$$\omega_i = \omega_r = \omega_t \ . \tag{4.44}$$

Then the exponential factors cancel out, and we have:

$$\boldsymbol{E}_{0i}^{\perp} + \boldsymbol{E}_{0r}^{\perp} = \boldsymbol{E}_{0t}^{\perp} \ . \tag{4.45}$$

In addition, since the frequency of the reflected and transmitted waves is the same as that of the incident wave (4.44), Equation 4.42 gives

$$\frac{\boldsymbol{k}_i \cdot \boldsymbol{k}_i}{n_i^2} = \frac{\boldsymbol{k}_r \cdot \boldsymbol{k}_r}{n_i^2} = \frac{\boldsymbol{k}_t \cdot \boldsymbol{k}_t}{n_t^2} \ . \tag{4.46}$$

In a sense this shows how the relation $\boldsymbol{k} \cdot \boldsymbol{k} = k_0^2 n^2$ governs the propagation of a plane wave in homogeneous matter ($k_0 = \omega/c$ is the wave number in vacuum).

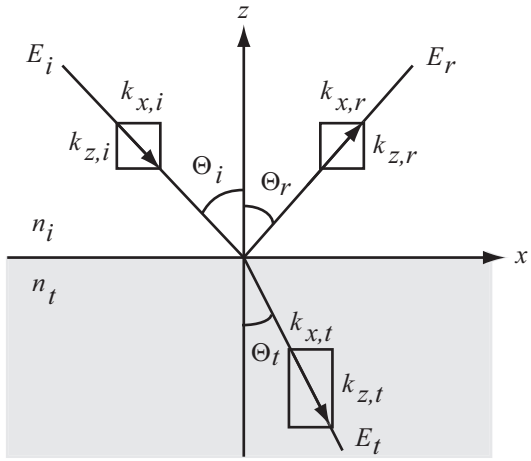


Figure 4.2: *A plane wave reflected and refracted on a surface $z = 0$ with normal in the direction of the z-axis. The xz-plane is the plane of incidence.*

Let us orient our coordinate system such that the tangent plane is the xy-plane and the plane of incidence is the xz-plane. Confer Figure 4.2. Then at the boundary, where $z = 0$, we have at the time $t = 0$

$$\boldsymbol{E}_{0i}^{\perp} e^{i(k_{x,i}x + k_{y,i}y)} + \boldsymbol{E}_{0r}^{\perp} e^{i(k_{x,r}x + k_{y,r}y)} = \boldsymbol{E}_{0t}^{\perp} e^{i(k_{x,t}x + k_{y,t}y)} \ ,$$

This must hold for all x and y on the boundary, thus

$$k_{x,i} = k_{x,r} = k_{x,t} \quad \text{and} \quad k_{y,i} = k_{y,r} = k_{y,t} \ . \tag{4.47}$$

Since $\boldsymbol{k}_i$ and the normal to the surface at the point of incidence span the plane of incidence, $\boldsymbol{k}_i$ is parallel to it and therefore has no y component, that is, $k_{y,i} = 0$. Then according to our new result (4.47), we have

$$k_{y,i} = k_{y,r} = k_{y,t} = 0 \ . \tag{4.48}$$

In other words the reflected and transmitted waves lie in the plane of incidence. Using Equation 4.46, we get

$$k_{x,i}^2 + k_{z,i}^2 = k_{x,r}^2 + k_{z,r}^2 \ ,$$

where the x components cancel out (cf. Equation 4.47), and the mathematical solution is $k_{z,r} = k_{z,i}$ or $k_{z,r} = -k_{z,i}$. The reflected wave was, however, defined to be propagating in the same medium as the incident wave, and if $k_{z,r}$ has the same sign as $k_{z,i}$, the reflected wave is moving across the boundary. Therefore the only solution that makes physical sense is

$$k_{z,r} = -k_{z,i} \ . \tag{4.49}$$

Equations 4.47, 4.48, and 4.49 summarise the law of reflection for plane waves: *The reflected wave lies in the plane of incidence, the (complex) angle of reflection is equal to the (complex) angle of incidence.*

The law of refraction is found in a similar way. From Equations 4.46 and 4.48 we have

$$\frac{\boldsymbol{k}_t \cdot \boldsymbol{k}_t}{n_t^2} = \frac{\boldsymbol{k}_i \cdot \boldsymbol{k}_i}{n_i^2} \ .$$

Dividing by $k_{x,t}^2 = k_{x,i}^2$ (the equality is from Equation 4.47) and juggling the terms around, the result is

$$n_i \sin\Theta_i = n_t \sin\Theta_t \ , \tag{4.50}$$

where $\sin\Theta_i = k_{x,i}/\sqrt{\boldsymbol{k}_i \cdot \boldsymbol{k}_i}$ is sine of the complex angle of incidence and $\sin\Theta_t = k_{x,t}/\sqrt{\boldsymbol{k}_t \cdot \boldsymbol{k}_t}$ is sine of the complex angle of refraction. We refer to this result as the generalised Snell's law . The law of refraction for plane waves is thus: *The refracted wave lies in the plane of incidence, the (complex) angle of refraction follows the generalised Snell's law* (4.50).

From the plane wave Maxwell equations (4.36) we have:

$$\boldsymbol{H}_0 = \frac{1}{\omega\mu}\boldsymbol{k} \times \boldsymbol{E}_0 \ .$$

Considering the x component of $\boldsymbol{H}_0$, and seing that $\boldsymbol{H}$ is also required to be continuous across the boundary, we have

$$-\frac{k_{z,i}}{\omega_i\mu_i}\boldsymbol{E}_{0i}^{\perp} - \frac{k_{z,r}}{\omega_r\mu_i}\boldsymbol{E}_{0r}^{\perp} = -\frac{k_{z,t}}{\omega_t\mu_t}\boldsymbol{E}_{0t}^{\perp} \ .$$

Recalling that the frequencies are equal (cf. Equation 4.44) and using the law of reflection $k_{z,r} = -k_{z,i}$, another way to write this is

$$k_{z,i}\boldsymbol{E}_{0i}^{\perp} - k_{z,i}\boldsymbol{E}_{0r}^{\perp} = k_{z,t}\frac{\mu_i}{\mu_t}\boldsymbol{E}_{0t}^{\perp} \ .$$

If we neglect the relative difference in permeability between the materials and insert our earlier result (4.45), this equation leads to

$$\boldsymbol{E}_{0r}^{\perp} = \frac{k_{z,i} - k_{z,t}}{k_{z,i} + k_{z,t}}\boldsymbol{E}_{0i}^{\perp} \ .$$

Dividing through by $\sqrt{\boldsymbol{k}_i \cdot \boldsymbol{k}_i}\boldsymbol{E}_{0i}^{\perp}$ and again using Equation 4.46, we obtain the Fresnel equation for the reflection of the $\perp$-polarised component of the light:

$$\tilde{r}_{\perp} = \frac{\boldsymbol{E}_{0r}^{\perp}}{\boldsymbol{E}_{0i}^{\perp}} = \frac{n_i\cos\Theta_i - n_t\cos\Theta_t}{n_i\cos\Theta_i + n_t\cos\Theta_t} \ , \tag{4.51}$$

where $\cos\Theta_i = k_{z,i}/\sqrt{\boldsymbol{k}_i \cdot \boldsymbol{k}_i}$ is cosine of the complex angle of incidence and $\cos\Theta_t = k_{z,t}/\sqrt{\boldsymbol{k}_t \cdot \boldsymbol{k}_t}$ is cosine of the complex angle of refraction. These complex angles that we have now mentioned a few times, do not have the same simple geometrical interpretation as real angles. Nevertheless, they are still useful because cosine of a complex angle is the dot product of two normalised complex vectors.

The $\parallel$-polarised component is obtained in a similar fashion, but in this case the electric field vector is no longer parallel to the tangential plane. To use the condition which says that the tangential component of the electric vector is continuous across the boundary, we have to project the $\parallel$-polarised component to the tangential plane using cosine of the complex angles, such that

$$\boldsymbol{E}_{\parallel i}\cos\Theta_i + \boldsymbol{E}_{\parallel r}\cos\Theta_r = \boldsymbol{E}_{\parallel t}\cos\Theta_t \ .$$

With this relation, and again requiring that $\boldsymbol{H}_0$ is continuous across the boundary, we obtain

$$\tilde{r}_\parallel = \frac{\boldsymbol{E}_{0r}^{\parallel}}{\boldsymbol{E}_{0i}^{\parallel}} = \frac{n_t\cos\Theta_i - n_i\cos\Theta_t}{n_t\cos\Theta_i + n_i\cos\Theta_t} \ . \tag{4.52}$$

The Fresnel equations (4.51,4.52) describe amplitude ratios, but often we are only interested in the flow of energy. To translate the amplitude ratios into energy ratios (reflectances), we square the absolute values [Born and Wolf 1999] such that

$$R_\perp = |\tilde{r}_\perp|^2 \quad \text{and} \quad R_\parallel = |\tilde{r}_\parallel|^2 \ .$$

The reflected $\perp$-polarised light is then the amount of incident $\perp$-polarised light times the reflectance $R_\perp$, and the amount of reflected $\parallel$-polarised light is the amount of incident $\parallel$-polarised light times the reflectance $R_\parallel$. For unpolarised light, the total reflectance is

$$R = \frac{1}{2}(R_\perp + R_\parallel) \ .$$

The transmittances are one minus the reflectances:

$$T_\perp = 1 - R_\perp \quad , \quad T_\parallel = 1 - R_\parallel \quad , \quad T = 1 - R \ .$$

The Fresnel equations illustrate that light may become polarised upon reflection. Polarisation was, however, used mostly as a mathematical convenience in the derivation of the Fresnel equations. Maxwell's equations do not give us any reason why polarisation is needed to model light. We can represent it in the wave theory, but we cannot explain why it is of any physical consequence. The photon spin is the reason why polarisation changes the properties of light. In

Chapter 3 we saw that photons are spin one particles which cannot exist in the rest state. Therefore only two angular momenta ($-\hbar$ and $\hbar$) are possible for photons. This means that two polarisation components (like the $\perp$ and $\parallel$ components chosen here) are appropriate.

This chapter has introduced several different ways to model electromagnetic radiation. The ways have been of diminishing exactitude. We have used more and more simplifications in order to describe a more and more macroscopic case. Let us briefly explore at what levels the material in this chapter allows us to construct rendering algorithms. As the most exact way of rendering using Maxwell's equations, we should use the general solution (4.16–4.17) for the microscopic equations (4.1–4.4). This solution involves only the simplifications that we discussed in the previous chapter as compared to the quantum theory. To succeed with such a renderer, we would have to model materials at an atomic level. It might be possible to model materials at a slightly more macroscopic level using the dipole approximation. We can derive current and charge densities from such dipoles and use them for the integrations that find the vector and scalar potentials of the field. One way to accomplish this integration is using molecular dynamics [Rapaport 2004]. If we compute the vector and scalar potentials for every patch on the image plane (every pixel), we can find the field vectors and with those we can find the Poynting vector leading to the colour values that we need. Of course we would need to evaluate the field at an appropriate number of wavelengths distributed throughout the visible part of the spectrum.

Taking one more step up the ladder towards a more feasible way of rendering realistic images, we arrived at the macroscopic Maxwell equations (4.20–4.23). The simplification was that we introduced two phenomenological vectors (the polarisation and magnetisation vectors) in order to represent materials at a more macroscopic level. Unfortunately we did not find an easier solution for the general version of the macroscopic Maxwell equations. Mostly we used them to move on to the time-harmonic Maxwell equations (4.30–4.33). The simplification at this point was that we introduced the isotropic material equations (4.27–4.29). Beside assuming that the materials are isotropic these equations also involve the simplifying assumptions that the permittivity, permeability, and conductivity of the materials are proportional to the field vectors $\boldsymbol{E}$ and $\boldsymbol{H}$. This is certainly not true in general, but the class of materials for which it is true is happily rather large. Unfortunately we did not find a simple general solution for the time-harmonic Maxwell equations either. To find that, we would need something more general than plane waves. When we move to geometrical optics in the next chapter, we will be able to model a more general type of wave at the cost of making assumptions about the wavelength. In this chapter we had to assume that the materials are also homogeneous in order to fit a plane wave solution to the time-harmonic equations.

It is certainly possible to construct a rendering algorithm based on the plane wave solution (4.34) of Maxwell's equations. With the additional simplification of using scalar waves this was indeed the solution employed by Moravec [1981] in his wave-theoretical rendering scheme. I do not know if an implementation of Moravec's algorithm has been attempted on modern hardware, but I suppose reasonable results would be obtained today in comparison to the less successful results in the original paper. Moravec's approach is to model waves by sweeping a 2D array of complex values over an entire scene and storing new sources at points where light is reflected. The reflected light is then propagated in the following sweep going in the opposite direction. The disadvantage of this approach is that the scene must be modelled by finely sliced "object planes" parallel to the image plane. Preferably there should be a plane for every half wavelength. This is a highly impractical and very storage intensive way of representing objects. Extraction of the planes from an implicit surface representation might be the way to go.

Even if we do not want to construct a wave-based rendering algorithm, the electromagnetic field theory is very useful for describing the interaction of light and matter in a more detailed way than would be possible if we did not know about it. The Fresnel equations are a good example of results that we cannot derive without employing some wave theory. So this is the crux of the matter: we need the wave theory (and in some cases even the quantum theory) to describe the interaction of light and matter when the simpler rendering methods fail to do so in sufficient detail. In this chapter we introduced the complex index of refraction at a phenomenological level. To predict theoretical values for the index of refraction, we would have to resort to a quantum description of matter. Similarly, we have to resort to the electromagnetic field theory to predict the phenomenological quantities which we use to describe the scattering of light in conventional rendering algorithms.

In the following chapters, we will move to the theories used for rendering in graphics today. You will see where the phenomenological description of scattering comes in, and a scheme for theoretical prediction of phenomenological scattering properties is presented (using Maxwell's equations) in Part II. First let us explore the most advanced ray theory available. It is often referred to as *geometrical optics*, and it is derived from the theory described in this chapter.

Chapter 5

Geometrical Optics

suddenly there fell across my path a glowing beam of sunshine that lighted up the grass before me. I stopped to see how the green blades danced in its light, how the sunshine fell down the sloping bank across the stream below. Whirring insects seemed to be suddenly born in its beam. The stream flowed more gayly, the flowers on its brim were richer in color.

Anonymous author of *Sunshine*
– in *The Atlantic Monthly*, Vol. 6, No. 38, pp. 657–667, 1860

The conventional algorithms for rendering realistic images use a ray theory of light. The goal of this chapter is to derive a ray theory which models the electromagnetic field as faithfully as possible. The rays which best represent electromagnetic waves, are relatively straightforward to find if the waves are assumed to be plane and moving in a non-absorbing, homogeneous dielectric. In fact it was shown by Sommerfeld and Runge [1911] that waves of this kind are *equivalent* to rays of light for $\lambda \to 0$, where λ denotes the wavelength in vacuum. Conferring the previous chapter, we realise that plane waves are very restricted. We would like our ray theory to handle a more general case. In particular we would like to be able to handle all kinds of isotropic materials. This means that we have to solve the time-harmonic Maxwell equations (4.30–4.33) rather than the plane wave Maxwell equations (4.35–4.38). Non-absorbing, homogeneous dielectrics is really a quite limited set of materials. By solving the more general set of equations, we also allow absorbing materials, heterogeneous (i.e. inhomogeneous) materials, and materials which are not necessarily dielectrics. The remaining restriction on the materials is that they are isotropic. The fundamen-

tal assumption of geometrical optics is that we can neglect wavelength ($\lambda \to 0$). This is often a good assumption in the visible part of the spectrum, where wavelengths range from 380 nm to 780 nm, but wave phenomena such as diffraction will not be captured in geometrical optics.

5.1 The Eikonal Equation

After application of some vector calculus to the time-harmonic Maxwell equations (4.30–4.33), we obtain the following second order wave equations (cf. Appendix A.1)

$$\nabla^2 \boldsymbol{E}_c + \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_c) + \nabla(\boldsymbol{E}_c \cdot \nabla \ln \varepsilon) = -k_0^2 n^2 \boldsymbol{E}_c \tag{5.1}$$

$$\nabla^2 \boldsymbol{H}_c + \nabla \ln \varepsilon \times (\nabla \times \boldsymbol{H}_c) + \nabla(\boldsymbol{H}_c \cdot \nabla \ln \mu) = -k_0^2 n^2 \boldsymbol{H}_c \ , \tag{5.2}$$

where $k_0 = 2\pi/\lambda$ is the wave number in vacuum and n is the (complex) index of refraction. If the wave is in a homogeneous medium, both μ, ε, σ and n will be independent of position. Such a scenario significantly simplifies the wave equations since $\nabla \ln \mu = \nabla \ln \varepsilon = \mathbf{0}$. Taking notice that $-k_0^2 \boldsymbol{E}_c$ and $-k_0^2 \boldsymbol{H}_c$ are the second derivatives with respect to time of the electric and magnetic fields, one discovers that Equations 5.1 and 5.2 reduce to first order wave equations when the medium is homogeneous.

Let us try the following wave functions (which are more general than plane waves) as a solution for the second order wave equations (5.1–5.2):

$$\boldsymbol{E}_c(\boldsymbol{x}, t) = \boldsymbol{E}_0(\boldsymbol{x}) e^{-i(\omega t - \boldsymbol{k}(\boldsymbol{x}) \cdot \boldsymbol{x})} \tag{5.3}$$

$$\boldsymbol{H}_c(\boldsymbol{x}, t) = \boldsymbol{H}_0(\boldsymbol{x}) e^{-i(\omega t - \boldsymbol{k}(\boldsymbol{x}) \cdot \boldsymbol{x})} \ . \tag{5.4}$$

Note that the vectors $\boldsymbol{k}$, $\boldsymbol{E}_0$, and $\boldsymbol{H}_0$ may depend on the position in the medium. As is customary in geometrical optics, we introduce the *optical path*:

$$\mathscr{S}(\boldsymbol{x}) = k_0^{-1} \boldsymbol{k}(\boldsymbol{x}) \cdot \boldsymbol{x} \ ,$$

and with this we write the wave functions (5.3–5.4) as follows:

$$\boldsymbol{E}_c(\boldsymbol{x}) = \boldsymbol{E}_0(\boldsymbol{x}) e^{i k_0 \mathscr{S}(\boldsymbol{x})} e^{-i\omega t} \tag{5.5}$$

$$\boldsymbol{H}_c(\boldsymbol{x}) = \boldsymbol{E}_0(\boldsymbol{x}) e^{i k_0 \mathscr{S}(\boldsymbol{x})} e^{-i\omega t} \ . \tag{5.6}$$

After insertion of the trial solutions (5.5–5.6) in the second order wave equations (5.1–5.2), the time exponentials cancel out. Application of some vector calculus

leads to equations of three terms summing to zero (see Appendix A.2). For the electric field we get

$$\left((\nabla\mathscr{S})^2 - n^2\right)\boldsymbol{E}_0 + (ik_0)^{-1}\boldsymbol{L}(\boldsymbol{E}_0, \mathscr{S}, \varepsilon, \mu) + (ik_0)^{-2}\boldsymbol{M}(\boldsymbol{E}_0, \varepsilon, \mu) = \boldsymbol{0} \ .$$

For large k_0 the second and third terms vanish. The fundamental assumption of geometrical optics ($\lambda \to 0$) corresponds to saying that $k_0 = 2\pi/\lambda$ is assumed very large. Thus the second and third terms definitely vanish in geometrical optics. It follows from the remaining term that

$$(\nabla\mathscr{S})^2 = \left(\frac{\partial\mathscr{S}}{\partial x}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial y}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial z}\right)^2 = n^2 \ . \tag{5.7}$$

This relation is known as the *eikonal equation*, and it is the basic equation of geometrical optics. It is the condition that must hold for our trial solution (5.3–5.4) to be a valid solution of Maxwell's equations. Details of the derivation is provided in Appendix A.2. Validity of the eikonal equation has previously been shown by Born and Wolf [1999] for heterogeneous media with real-valued indices of refraction, and by Bell [1967] for homogeneous media with complex indices of refraction. Epstein [1930] derived the eikonal equation in as general a form as we do, but he derived it from the first order wave equation $\left(\nabla^2\boldsymbol{E} + k_0^2 n^2 \boldsymbol{E} = 0\right)$ rather than Maxwell's equations. The details of the derivation are thus included because they show why the eikonal equation is also valid in a heterogeneous and possibly absorbing medium. Put in a different way, it is shown in Appendix A.2 that the eikonal equation (5.7) is valid even when n and $\mathscr{S}$ are complex functions of the position $\boldsymbol{x}$ in the medium.

5.2 The Direction of Energy Propagation

As we have indicated in previous chapters, we are not really interested in tracing the wave fronts. The important thing to trace in a rendering context is the energy. Therefore we want the rays of light to follow the direction of energy flow in the field. This direction is given by Poynting's vector (cf. Equation 4.5):

$$\boldsymbol{S} = \varepsilon_0 c^2 \mathrm{Re}(\boldsymbol{E}_c) \times \mathrm{Re}(\boldsymbol{B}_c) = \varepsilon_0 c^2 \mu \mathrm{Re}(\boldsymbol{E}_c) \times \mathrm{Re}(\boldsymbol{H}_c) \ .$$

Observe that Poynting's vector is orthogonal to the real parts of both the electric and the magnetic vectors. The gradient of the optical path

$$\nabla\mathscr{S}(\boldsymbol{x}) = k_0^{-1}(\boldsymbol{k}(\boldsymbol{x}) + (\boldsymbol{x}\cdot\nabla)\boldsymbol{k}(\boldsymbol{x})) \tag{5.8}$$

has similar properties in some special cases. For homogeneous *waves* (which have $\boldsymbol{k}'$ and $\boldsymbol{k}''$ parallel) the real part of $\nabla\mathscr{S}$ is also the direction of Poynting's

vector (this is shown later in this chapter). For homogeneous *media* $\nabla \boldsymbol{k} = \boldsymbol{0}$, and then it follows from Equation 5.8 that $\nabla \mathscr{S}$ and $\boldsymbol{k}$ are parallel. This means that the waves are plane, and $\nabla \mathscr{S}$ corresponds to the wave vector, but the directions of the real part $\boldsymbol{k}'$ and the imaginary part $\boldsymbol{k}''$ are not necessarily parallel. Thus even for homogeneous media the gradient of the optical path $\nabla \mathscr{S}$ is not necessarily the direction of Poynting's vector. When the real and imaginary parts of the wave vector are not parallel, the wave is referred to as inhomogeneous. Except for a few special cases, waves are always inhomogeneous when propagating in absorbing media. This is the case for both homogeneous and heterogeneous absorbing media (and sometimes, but rarely, also for non-absorbing media). Therefore inhomogeneous waves are an important type of light which we ought to be able to handle in geometrical optics.

Since the direction of damping $\boldsymbol{k}''$ and the direction normal to the surface of constant phase $\boldsymbol{k}'$ are not parallel for inhomogeneous waves, the damping causes Poynting's vector to start oscillating. These oscillations are so rapid that the instantaneous direction of Poynting's vector does not give any useful information about the general progress of energy in the field. Hence, we have to come up with a different way of finding the direction of the energy flow. The generally accepted approach is to consider the time average of Poynting's vector $\boldsymbol{S}_{\text{avg}}$ over a period of oscillation $T = 2\pi/\omega$:

$$\boldsymbol{S}_{\text{avg}} = \frac{\varepsilon_0 c^2 \mu}{T} \int_0^T \text{Re}(\boldsymbol{E}_c) \times \text{Re}(\boldsymbol{H}_c)\, \text{d}t \ . \tag{5.9}$$

Considering our solution (5.3–5.4), we write the electric and magnetic vectors as follows:

$$\begin{aligned} \boldsymbol{E}_c(\boldsymbol{x}, t) &= (\boldsymbol{E}_0'(\boldsymbol{x}) + i\boldsymbol{E}_0''(\boldsymbol{x}))e^{-i(\omega t - \boldsymbol{k}'(\boldsymbol{x})\cdot\boldsymbol{x})} e^{-\boldsymbol{k}''(\boldsymbol{x})\cdot\boldsymbol{x}} \\ \boldsymbol{H}_c(\boldsymbol{x}, t) &= (\boldsymbol{H}_0'(\boldsymbol{x}) + i\boldsymbol{H}_0''(\boldsymbol{x}))e^{-i(\omega t - \boldsymbol{k}'(\boldsymbol{x})\cdot\boldsymbol{x})} e^{-\boldsymbol{k}''(\boldsymbol{x})\cdot\boldsymbol{x}} \ . \end{aligned}$$

Setting $\theta(\boldsymbol{x}, t) = \omega t - \boldsymbol{k}'(\boldsymbol{x}) \cdot \boldsymbol{x}$, we get the following real parts

$$\text{Re}(\boldsymbol{E}_c) = (\boldsymbol{E}_0' \cos\theta - \boldsymbol{E}_0'' \sin\theta) e^{-\boldsymbol{k}''\cdot\boldsymbol{x}} \tag{5.10}$$

$$\text{Re}(\boldsymbol{H}_t) = (\boldsymbol{H}_0' \cos\theta - \boldsymbol{H}_0'' \sin\theta) e^{-\boldsymbol{k}''\cdot\boldsymbol{x}} \ . \tag{5.11}$$

These real parts still have the same period of oscillation, and since the integral of a sine or a cosine over a period of oscillation is zero, insertion in Equation 5.9 gives (see Appendix A.3)

$$\boldsymbol{S}_{\text{avg}} = \frac{\varepsilon_0 c^2 \mu}{2} (\boldsymbol{E}_0' \times \boldsymbol{H}_0' + \boldsymbol{E}_0'' \times \boldsymbol{H}_0'') e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ . \tag{5.12}$$

To have an easier way of modelling inhomogeneous waves, we resolve the wave functions into two independent components. One is the *transverse electric* (TE)

component, and it has the feature that $\boldsymbol{E}_0 = \boldsymbol{E}_0'$. The other component is the *transverse magnetic* (TM) , which has $\boldsymbol{H}_0 = \boldsymbol{H}_0'$. This choice of components imposes no restrictions on the types of waves that we are able to model. Let us determine the direction of the time averaged energy flux $\boldsymbol{S}_{\text{avg}}$ in each component. Using the fundamental assumption in geometrical optics ($\lambda \to 0$), and the properties $\boldsymbol{E}_0 = \boldsymbol{E}_0'$ and $\boldsymbol{H}_0 = \boldsymbol{H}_0'$ for the TE wave and TM wave respectively, we find (see again Appendix A.3)

$$\boldsymbol{S}_{\text{avg,TE}} = \frac{\varepsilon_0 c}{2}|\boldsymbol{E}_0|^2 \text{Re}(\nabla\mathscr{S})e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \tag{5.13}$$

$$\boldsymbol{S}_{\text{avg,TM}} = \frac{\varepsilon_0 c^3\mu^2}{2}|\boldsymbol{H}_0|^2 \text{Re}\left(\frac{\nabla\mathscr{S}}{n^2}\right) e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ . \tag{5.14}$$

This shows that the energy flux carried by a TE wave follows the direction of the real part of $\nabla\mathscr{S}$. A TM wave, however, follows a slightly different path in which both the real and imaginary parts of $\nabla\mathscr{S}$ are necessary. Since the eikonal equation (5.7) is valid for complex $\nabla\mathscr{S}$, its solution provide sufficient information for us to find the direction of rays representing TE and TM waves. Therefore we will use the eikonal equation in the following to derive the path of such rays.

First we would like to express the two Poynting vector components (5.13–5.14) using only the optical path $\mathscr{S}$, the material properties, and some initial energy $|\boldsymbol{S}_0|$. Otherwise we cannot make a ray theory of light which is independent of the electromagnetic field vectors (and that is what we are aiming at in this chapter). To begin with, let us find out how $|\boldsymbol{H}_0|$ relates to $|\boldsymbol{E}_0|$. We will find this relation using a Equation A.13 which was derived from Maxwell's equations in Appendix A.3. The squared absolute value of a complex vector is defined by

$$|\boldsymbol{v}|^2 = \boldsymbol{v}\cdot\boldsymbol{v}^* \ ,$$

where the asterisk * denotes the complex conjugate. We take the squared magnitude on both sides of Equation A.13, and get

$$|\boldsymbol{H}_0|^2 = \frac{1}{c^2\mu^2}|\nabla\mathscr{S}\times\boldsymbol{E}_0|^2 = \frac{1}{c^2\mu^2}(|\nabla\mathscr{S}|^2|\boldsymbol{E}_0|^2 - |\nabla\mathscr{S}\cdot\boldsymbol{E}_0|^2) \ .$$

To find the magnitude of the gradient of the optical path $|\nabla\mathscr{S}|$, we take the absolute value on both sides of the eikonal equation (5.7). The result is

$$|\nabla\mathscr{S}|^2 = |n|^2 \ .$$

With this result, and the condition $\boldsymbol{E}_0\cdot\nabla\mathscr{S} = 0$ (cf. Equation A.14), we arrive at the relation

$$|\boldsymbol{H}_0|^2 = \frac{|n|^2}{c^2\mu^2}|\boldsymbol{E}_0|^2 \ . \tag{5.15}$$

Using this relation between $|\boldsymbol{E}_0|$ and $|\boldsymbol{H}_0|$ in Equations 5.13 and 5.14, we get the formulae for the two components of the Poynting vector on the form that we were looking for:

$$\boldsymbol{S}_{\text{avg,TE}} = |\boldsymbol{S}_0|\frac{1}{n'}\text{Re}(\nabla\mathscr{S})e^{-2k_0\text{Im}(\mathscr{S})} \tag{5.16}$$

$$\boldsymbol{S}_{\text{avg,TM}} = |\boldsymbol{S}_0|\frac{|n|^2}{n'}\text{Re}(\nabla\mathscr{S}/n^2)e^{-2k_0\text{Im}(\mathscr{S})} \ , \tag{5.17}$$

where the initial energy is

$$|\boldsymbol{S}_0| = \frac{\varepsilon_0 c}{2}n'|\boldsymbol{E}_0|^2 \ .$$

A ray of energy $|\boldsymbol{S}_0|$ could be any ray of light. This means that we can use the equations (5.16–5.17) to split a ray of unpolarised light into it's TE and TM components, and after the split the total energy of the ray will be half the sum of the two components:

$$\boldsymbol{S}_{\text{avg}} = \frac{1}{2}(\boldsymbol{S}_{\text{avg,TE}} + \boldsymbol{S}_{\text{avg,TM}}) \ .$$

One should take note of Equations 5.16 and 5.17 (I have seen them nowhere else in the literature, the closest are equations similar to 5.13 and 5.14 derived for homogeneous media by Bell [1967, p. 8]. Bell does not find them useful because his purpose is not to trace rays, but rather to calculate formulae for measuring optical properties). As we will see shortly, it is very useful for tracing rays in absorbing media that we are able to split a ray in its TE and TM components.

A homogeneous wave ($\text{Re}(\nabla\mathscr{S})$ and $\text{Im}(\nabla\mathscr{S})$ parallel) moves in the direction normal to the surface of constant phase, and the direction of a light ray representing a homogeneous wave is found directly by solving the eikonal equation. Making a ray that represents an inhomogeneous wave is unfortunately not so simple. We cannot merely add the directions of the two Poynting vector components to get the direction of the energy flux carried by an inhomogeneous wave. Using a relation (A.13) derived in Appendix A.3 with the expression for the time-averaged Poynting vector (5.12), the true direction (written in terms of the electric field vector) is

$$\boldsymbol{S}_{\text{avg}} = \frac{\varepsilon_0 c}{2}\left(|\boldsymbol{E}_0|^2\text{Re}(\nabla\mathscr{S}) - \text{Re}(\boldsymbol{E}_0(\boldsymbol{E}_0^*\cdot\nabla\mathscr{S}))\right)e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ . \tag{5.18}$$

There seems to be no way to represent this direction without reference to the electric (or the magnetic) field vector. This means that we cannot correctly represent inhomogeneous waves by rays of light. One solution is to trace the wave front using complex vectors and angles. This is called *complex ray tracing*

[kravtsov 1967; Bennett 1974; Chapman et al. 1999], but it is not really a ray theory of light. To model inhomogeneous waves in a ray theory of light, we have to make an approximation. The traditional approximation is simply to trace $\mathrm{Re}(\nabla\mathscr{S})$ and neglect the change in direction caused by the absorption. This is a good approximation for weakly absorbing materials.

As an alternative approximation for modelling inhomogeneous waves in geometrical optics, my proposal is to trace two rays: one for the TE component and one for the TM component. This would capture the propagation of light in absorbing media more faithfully. The energy flux is split in two components propagating in slightly different directions. These are directions that we are able to find without keeping track of the electromagnetic field vectors (using Equations 5.16 and 5.17). The total energy flux of the represented wave is then a weighted sum of the magnitudes of the two components. The underlying assumption is that the medium changes smoothly such that the weighted energy flow in two slightly different directions is a good representative of the flow in the correct direction (5.18). The correct direction is in-between the directions of the two components.

How do we weight the TE and TM components in a ray tracing? Luckily there is nothing preventing us from choosing to identify the TE wave with the $\perp$-polarised wave and the TM wave with the $\parallel$-polarised wave described in the previous chapter. In other words we choose to represent a wave of light by two components: A transverse electric component with the electric vector perpendicular to the plane of incidence (TE,$\perp$) and a transverse magnetic component with the electic vector parallel to the plane of incidence (TM,$\parallel$). This makes the answer much easier: If unpolarised light is emitted in an absorbing medium, we weight the TE and TM components by one half each:

$$|\boldsymbol{S}_{\mathrm{avg}}|^{\mathrm{unpolarised}} = \frac{1}{2}(|\boldsymbol{S}_{\mathrm{avg,TE}}| + |\boldsymbol{S}_{\mathrm{avg,TM}}|) \ .$$

If light is not emitted in an absorbing medium, but refracts into an absorbing medium, we weight the TE component by the amount of $\perp$-polarised transmitted light and the TM component by the amount of $\parallel$-polarised transmitted light:

$$|\boldsymbol{S}_{\mathrm{avg}}|^{\mathrm{transmitted}} = T_{\perp}|\boldsymbol{S}_{\mathrm{avg,TE}}| + T_{\parallel}|\boldsymbol{S}_{\mathrm{avg,TM}}| \ .$$

5.3 Tracing Rays of Light

Knowing how to find the energy flux associated with a ray, the next thing we need to do, is to find the path that a ray follows. In other words we need to

solve the eikonal equation (5.7). Let us find a general solution for the eikonal equation using the Jacobi method. Consider the gradient of the optical path:

$$\nabla \mathscr{S} = \left(\frac{\partial \mathscr{S}}{\partial x}, \frac{\partial \mathscr{S}}{\partial y}, \frac{\partial \mathscr{S}}{\partial z} \right) = (p, q, r) = \boldsymbol{p} \ .$$

This way of writing it illustrates that we can write the eikonal equation as a partial differential equation of the form

$$H(p, q, r, x, y, z) = 0 \ , \tag{5.19}$$

where H corresponds to the Hamiltonian function in dynamics. In the case of the eikonal equation, H is given by

$$H(p, q, r, x, y, z) = p^2 + q^2 + r^2 - n^2 \ , \tag{5.20}$$

where both p, q, r, and n depend on $\boldsymbol{x} = (x, y, z)$.

Now we introduce the parametric equations $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$ to denote a ray. Then the derivatives with respect to τ are given by

$$\frac{dx}{d\tau} = \alpha \frac{\partial \mathscr{S}}{\partial x} \ , \quad \frac{dy}{d\tau} = \alpha \frac{\partial \mathscr{S}}{\partial y} \ , \quad \frac{dz}{d\tau} = \alpha \frac{\partial \mathscr{S}}{\partial z} \ , \tag{5.21}$$

where α can be chosen arbitrarily, since it has no influence on the geometrical path of the ray. It only has influence on the unit of arc length along the ray [Kline and Kay 1965]. Suppose we choose $\alpha = 2$, then by the definition of H, we may write the derivatives above as follows:

$$\frac{dx}{d\tau} = \frac{\partial H}{\partial p} \ , \quad \frac{dy}{d\tau} = \frac{\partial H}{\partial q} \ , \quad \frac{dz}{d\tau} = \frac{\partial H}{\partial r} \ . \tag{5.22}$$

Since p, q, and r are also functions of τ, these equations are not adequate for finding the path of the rays. But now all we need to find are the derivatives of p, q, and r with respect to τ.

Observe that x does not appear explicitly in Equation 5.20. Consequently we can express H as a function of the explicit intermediate variables p, q, and r, and the independent variable x. Then according to the chain rule

$$\frac{dH}{dx} = \frac{\partial H}{\partial x} + \frac{\partial H}{\partial p}\frac{dp}{dx} + \frac{\partial H}{\partial q}\frac{dq}{dx} + \frac{\partial H}{\partial q}\frac{dq}{dx} \ .$$

By virtue of Equations 5.22, we obtain

$$\frac{dH}{dx} = \frac{\partial H}{\partial x} + \frac{dp}{d\tau} + \frac{dy}{dx}\frac{dq}{d\tau} + \frac{dz}{dx}\frac{dr}{d\tau} \ .$$

According to the condition (5.19), $dH/dx = 0$. Furthermore y and z are independent of x, thus

$$\frac{dp}{d\tau} = -\frac{\partial H}{\partial x} \ , \tag{5.23}$$

and using the same line of arguments only differentiating H with respect to y and z, we get

$$\frac{dq}{d\tau} = -\frac{\partial H}{\partial y} \ , \quad \frac{dr}{d\tau} = -\frac{\partial H}{\partial z} \ . \tag{5.24}$$

Let $\mathscr{S}(x, y, z, \alpha_1, \alpha_2)$ be a complete integral of Equation 5.19, where α_1, α_2 is a set of arbitrary independent parameters. Then the theorem of Jacobi says that the six ordinary differential equations (5.22-5.24) has the solution given by the four-parameter manifold of curves represented by the equations [Kline and Kay 1965]

$$\frac{\partial \mathscr{S}}{\partial \alpha_1} = \beta_1 \quad \text{and} \quad \frac{\partial \mathscr{S}}{\partial \alpha_2} = \beta_2 \ , \tag{5.25}$$

where β_1, β_2 is another set of arbitrary parameters. This is all the fundamental theory we need to find $\nabla\mathscr{S}$ in isotropic, heterogeneous media. Recall that $\nabla\mathscr{S}$ is not the direction of rays of light in the media. The gradient of the optical path $\nabla\mathscr{S}$ is a complex-valued vector giving us the means to find the direction of rays of light using Equations 5.16 and 5.17.

The solution is a little abstract, so let us construct an example. Take an isotropic medium in which the index of refraction n depends only on the coordinate z. In this case the eikonal equation (5.7) attains the form

$$\left(\frac{\partial \mathscr{S}}{\partial x}\right)^2 + \left(\frac{\partial \mathscr{S}}{\partial y}\right)^2 + \left(\frac{\partial \mathscr{S}}{\partial z}\right)^2 = n^2(z) \ .$$

Using separation of variables, the complete integral of this equation is given by

$$\mathscr{S} = ax + by + \int \sqrt{n^2(z) - a^2 - b^2}\, \mathrm{d}z \ ,$$

where a and b are arbitrary parameters. With this solution, we let $\alpha_1 = a$ and $\alpha_2 = b$, then application of Equation 5.25 results in

$$x - \alpha_1 \int \frac{1}{\sqrt{n^2(z) - \alpha_1^2 - \alpha_2^2}}\, \mathrm{d}z \ = \ \beta_1 \tag{5.26}$$

$$y - \alpha_2 \int \frac{1}{\sqrt{n^2(z) - \alpha_1^2 - \alpha_2^2}}\, \mathrm{d}z \ = \ \beta_2 \ . \tag{5.27}$$

Through variation of the four parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2)$, these two equations specify the path of all possible complex paths in the medium, and a complex path specifies the direction of $\nabla\mathscr{S}$.

To be specific, we introduce a surface of discontinuity at $z = 0$, and say that a ray is incident from Medium 1 ($z > 0$) with index of refraction n_1 at an angle θ_1 with the z-axis. In other words, the ray has the initial condition

$$\mathscr{S} = n_1 \sin\theta_1 \, x + n_1 \cos\theta_1 \, z \ .$$

Then $\alpha_1 = n_1 \sin\theta_1$, $\alpha_2 = 0$ and we choose $\beta_1 = \beta_2 = 0$. According to the solution (5.26–5.27), we have

$$x = \int \frac{n_1 \sin\theta_1}{\sqrt{n_2^2(z) - (n_1 \sin\theta_1)^2}} \, dz \tag{5.28}$$

$$y = 0 \ , \tag{5.29}$$

where n_1 and n_2 are the (complex) refractive indices of Medium 1 and Medium 2, and $y = 0$ says that the path stays in the plane of incidence.

Finding the differential form of Equation 5.28 at the discontinuity, gives the (complex) angle of refraction Θ_2 into Medium 2 ($z < 0$). The result is

$$\tan\Theta_2 = \frac{dx}{dz} = \frac{n_1 \sin\theta_1}{\sqrt{n_2^2 - n_1^2 \sin^2\theta_1}} \ , \tag{5.30}$$

Some symbolic manipulation shows that this is equivalent to the generalised Snell's law (4.50), and if either Medium 1 or Medium 2 is absorbing, the resulting angle of refraction Θ_2 is probably complex. The direction of the refracted path is the gradient of the optical path $\nabla\mathscr{S}$, which is what we are looking for.

Let $\vec{n}$ denote a unit vector normal to the surface of discontinuity pointing into Medium 1 at the location $\boldsymbol{x}$. Suppose $\vec{t}$ is a unit vector tangent to the surface at $\boldsymbol{x}$. Then the direction of the refracted light is

$$\nabla\mathscr{S} = n_2 \sin\Theta_2 \, \vec{t} - n_2 \cos\Theta_2 \, \vec{n} \ . \tag{5.31}$$

Let $\vec{\omega}'$ denote the direction from which a ray of light is incident at $\boldsymbol{x}$. The component of $\vec{\omega}'$ which is perpendicular to the normal, is

$$\vec{\omega}'_\perp = (\vec{n} \cdot \vec{\omega}')\vec{n} - \vec{\omega}' \ ,$$

and we have

$$|\vec{\omega}'_\perp| = \sin\theta_1 \ .$$

An expression for the unit tangent vector $\vec{t}$ then follows:

$$\vec{t} = \frac{\vec{\omega}'_\perp}{|\vec{\omega}'_\perp|} = \frac{(\vec{n} \cdot \vec{\omega}')\vec{n} - \vec{\omega}'}{\sin\theta_1} \ . \tag{5.32}$$

Since we can assume that $\cos\Theta_2$ is in the first quadrant, another way to express $\cos\Theta_2$ is

$$\cos\Theta_2 = \sqrt{1-\sin^2\Theta_2} = \sqrt{1-(n_1/n_2)^2\sin^2\theta_1}\ , \tag{5.33}$$

where we have used the generalised Snell's law, or, equivalently, Equation 5.30 written in a different way.

Inserting from Equations 5.32 and 5.33 in Equation 5.31, and using the generalised Snell's law again, we get the usual formula for finding the direction of refracted light (cf. Glassner's [1995] Equation 11.57):

$$\nabla\mathscr{S} = n_1\left((\vec{n}\cdot\vec{\omega}')\vec{n}-\vec{\omega}'\right)+\vec{n}\sqrt{n_2^2-n_1^2(1-(\vec{n}\cdot\vec{\omega}')^2)}\ . \tag{5.34}$$

Our version does not give a vector of unit length and the indices of refraction are now complex numbers. The quantity we find with our slightly different formula (5.34) is the gradient of the optical path $\nabla\mathscr{S}$ (a complex-valued vector) instead of the direction of a refracted ray of light. The gradient of the optical path is what we need to find the directions of the two rays representing an inhomogeneous wave: We use $\mathrm{Re}(\nabla\mathscr{S})$ for the direction of the $\perp$-polarised component and $\mathrm{Re}(\nabla\mathscr{S}/n_2^2)$ for the direction of the $\parallel$-polarised component. These two components are treated as normal rays of light with energy given by $T_\perp|\boldsymbol{S}_{\mathrm{avg,TE}}|$ and $T_\parallel|\boldsymbol{S}_{\mathrm{avg,TM}}|$ respectively, where $T_\perp$ and $T_\parallel$ are the Fresnel transmittances discussed in Section 4.4. If the real and imaginary parts of $\nabla\mathscr{S}$ are parallel (and that includes when there is no imaginary part), the wave we are representing is homogeneous and then we need only one ray traced in the direction $\mathrm{Re}(\nabla\mathscr{S})$ to represent it.

5.4 Rendering Small Absorbing Particles

The rather general exposition of geometrical optics given in this chapter describes a ray theory of light which is valid for representing electromagnetic waves of very short wavelength in isotropic, heterogeneous, absorbing materials. Of course homogeneous materials and non-absorbing (or transparent) materials are special cases of our more general theory. With some geometrical representation of the media in a scene, a ray theory of light as the one presented here, is directly applicable as a rendering algorithm. To my knowledge, the eikonal equation was first employed in graphics by Stam and Languénou [1996]. Their mission was to render transparent, heterogeneous media. More recently Irhke et al. [2007] presented a very efficient way of rendering using the eikonal equation. They include absorption and scattering effects, but all at a simplified level. This work and other work on light transport in heterogeneous media [Gutierrez et al. 2005, for example] do not consider the effect of the imaginary part of the

refractive index. Glassner [1995] mentions the Fresnel equations for materials with complex index of refraction, but he does not include the imaginary part in his derivation of Snell's law. Hopefully this chapter sheds some light on the role of the imaginary part of the refractive index in the propagation of waves of light.

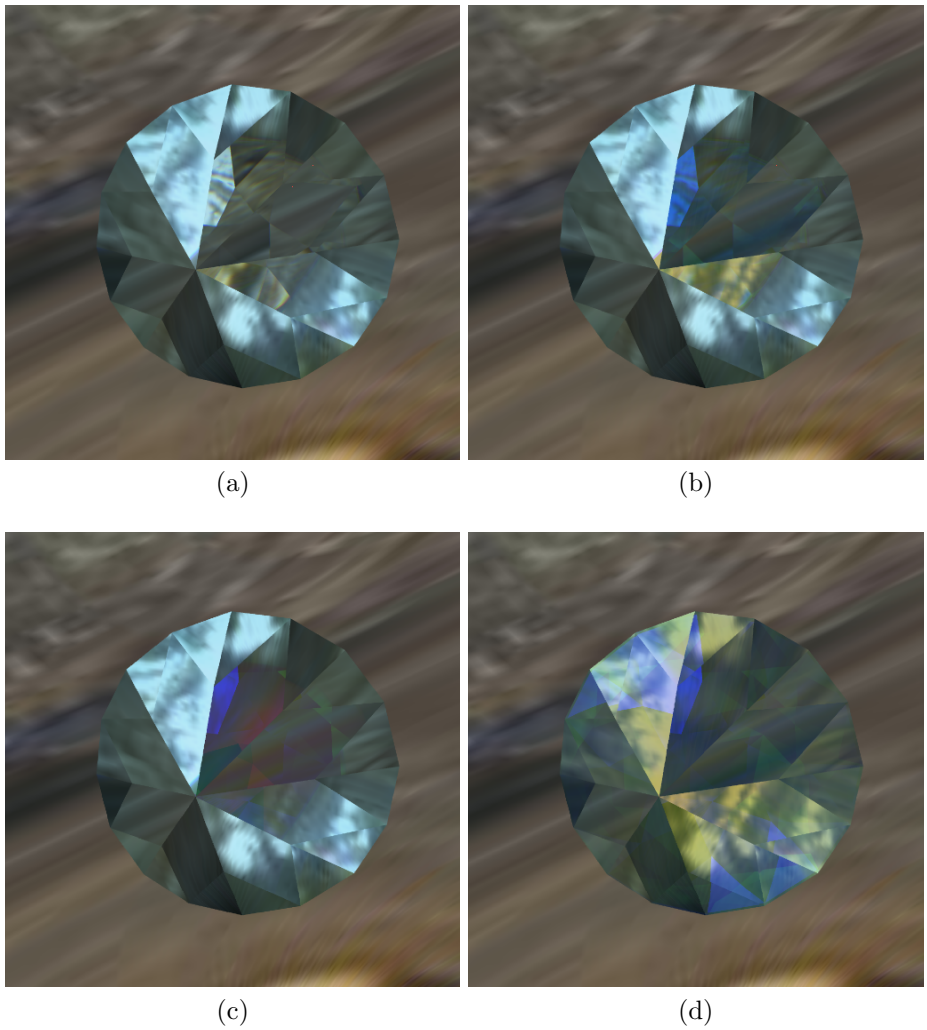
When light passes from a dense transparent medium to a sparse transparent medium, say from glass to air, there are angles of incidence for which the reflectance is unity. This happens when cosine of the complex angle of refraction $\cos\Theta_2$ becomes purely imaginary because then the Fresnel equations for reflectance (4.51,4.52) return 1. The smallest angle of total reflection is called the critical angle θ_c. From Equation 5.33 it follows that

$$\sin\theta_c = n_2/n_1 \qquad (n_1 \text{ and } n_2 \text{ real}) \ .$$

For these angles of incidence $\theta_1 \geq \theta_c$, that is, for purely imaginary $\cos\Theta_2$, we have the phenomenon called *total internal reflection*. An interesting consequence of absorption is that it eliminates total internal reflection. If either of the refractive indices (n_1 or n_2) is a complex number, there will always be some transmitted light.

The reason why the imaginary part of the refractive index has not been given much thought in graphics, is probably that it is often very small when the wavelength is short (this follows from the relation (4.43) between wavelength, absorption coefficient, and the imaginary part of the refractive index). If it is not very small, the absorption is very strong, and then we cannot make out the effect of refraction anyway. This is a very convincing argument why we do not need to worry about the imaginary part of the refractive index when computing the direction of a refracted ray. For unusual graphics applications it is, however, important to know about the effect of absorption on the refraction of light. If we were going to use graphics techniques for visualizing radio waves, the effect of the imaginary part of the refractive index would be very important. The wavelengths of radio waves easily range from meters to kilometers. To take an example of importance when we consider light, that is, electromagnetic waves with visible wavelengths, let us consider very small, strongly absorbing particles.

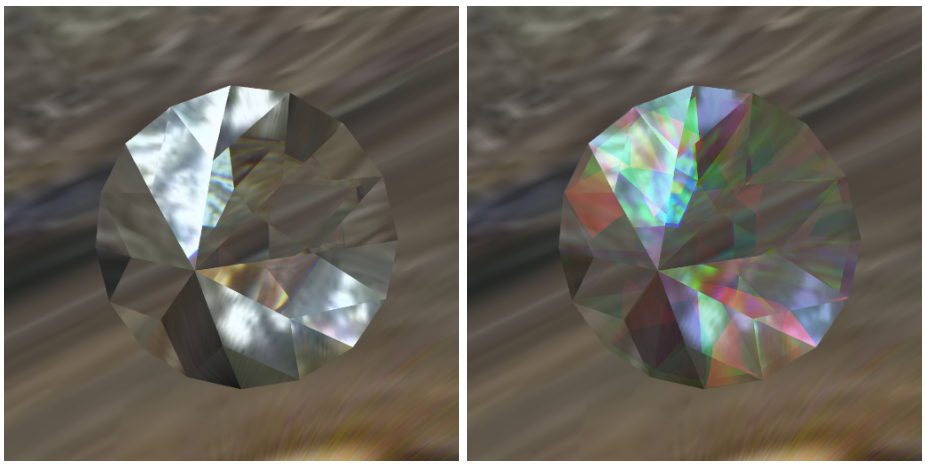
Figure 5.1a shows a standard ray tracing of diamond shaped particle. The particle is surrounded by a smeared out background which is modelled as if it were very far away. The size of the particle is around a nanometer and the absorption coefficient is around $10^9\,\mathrm{m}^{-1}$. This makes the imaginary part of the refractive index of the same scale as the real part (around 1). The spectrum used for the absorption is that of pure ice scaled by 10^9 such that it is of the same scale as the absorption of a blue pigment. The result of the standard ray tracing looks pretty much like a diamond with a little blue absorption in it. If we use the complex index of refraction when evaluating the Fresnel equations



(a) (b)

(c) (d)

Figure 5.1: *Renderings of a small (nanometer scale), strongly absorbing, diamond shaped particle. The absorption spectrum is that of pure ice scaled by* 10^9 *such that it corresponds to a blue pigment. The images in the top row (a,b) were rendered using standard ray tracing, the images in the bottom row (c,d) were rendered using the geometrical optics presented in this chapter. In the left column (a,c) the real part of the refractive index was used in the Fresnel equations, in the right column (b,d) the complex index of refraction was used.*



(a) (b)

Figure 5.2: *The same diamond shaped particle as in Figure 5.1. This time the absorption is the spectrum of ice scaled by* 10^8*. From left to right: (a) Standard ray tracing. (b) Rendering using the geometrical optics presented in this chapter. Both renderings use the complex index of refraction for the Fresnel equations (with or without does not make much difference in this case).*

in the standard ray tracer, the result is very different (Fig. 5.1b). The reason is that more background lighting is allowed to refract into the particle. What happens when we use the ray theory of light presented in this chapter? If we use it with only the real part of the refractive index in the Fresnel equations (Fig. 5.1c), the dispersive effects show themselves very clearly. They are caused by the influence of the absorption on the direction of the refracted light. The correct rendering using the complex index of refraction throughout, is presented in Figure 5.1d. To show that the dispersive effect diminishes as the absorption decreases, Figure 5.2 presents the effect when the absorption is 10 times less than in Figure 5.1. If the absorption is reduced by another factor 10 while the size of the particle is increased by a factor 100, the effect is only just discernable.

Tracing rays of light through small particles provides a simple way of simulating their light scattering properties. Later in this thesis (Part II) we will see that it is rather difficult to determine, mathematically, how very small particles scatter light. At least this is true if the particles are not approximately spherical. The geometrical optics presented in this chapter enable us to compute the scattering by particles of arbitrary geometry. The theory includes refraction effects caused by strong absorption, and we have demonstrated by example that it is very important to include these effects if we want to capture the scattering properties of small absorbing particles correctly.

Using a ray theory of light to compute the optical properties of small particles with arbitrary surface geometry, is (at least in graphics) a previously unexplored way of finding optical properties. The most similar approach I am aware of is the idea of having a *shell transport function* which captures the light transport in a representative lump of medium containing arbitrarily shaped particles [Moon et al. 2007]. Such a shell transport function is quite similar to the scattering properties of a medium. The difference is that it describes a scalable sphere containing the medium rather than the properties of the medium at a specific point. As for other work in graphics, the shell transport functions do not take the refraction effects caused by absorption into account. I believe that the idea of capturing optical properties using a ray tracing approach presented in this chapter is an interesting concept that should be investigated further in the future.

In this chapter we have arrived at a ray theory of light which is readily applicable in rendering. The simplification we used to obtain a ray theory of light from the macroscopic Maxwell equations, was to assume that wavelength is very small. This is a good assumption in most cases involving light because electromagnetic waves in the visible part of the spectrum have very short wavelengths. Using this assumption we showed how to trace light through isotropic materials. It should be mentioned that it is possible to generalise the theory we have presented such that it also allows anisotropic materials (see for example the work of Kline and Kay [1965]). Although we used the macroscopic Maxwell equations, the level at which we model materials is still very microscopic compared to conventional rendering techniques. We still need to model every particle that has an index of refraction which is different from the surroundings. In the next chapter we introduce macroscopic material properties to describe the scattering of light at a more macroscopic level. This leads us to radiative transfer theory which is the theory used for most realistic rendering.

Chapter 6

Radiative Transfer

In the beginning, in the dark, there was nothing but water, and Bumba was alone.

One day Bumba was in terrible pain. He retched and strained and vomited up the sun. After that light spread over everything. The heat of the sun dried up the water until the black edges of the world began to show.

from a creation story of the Boshongo People

Radiative transfer theory was introduced to graphics by Jim Blinn [1982]. Its purpose was, and still is today, to simulate the scattering of light that goes on beneath the surfaces of semi-transparent objects. We saw in the previous chapter how a ray theory of light is useful for simulating reflection and refraction in specular objects. The problem is that most objects are composed of millions of specular particles. There are far too many of them for us to trace rays through every single one. To simulate the scattering of light by a large number of small particles, Arthur Schuster [1905] introduced the *scattering coefficient*; a new macroscopic, or phenomenological, quantity to join the index of refraction which sums up the material properties used in the macroscopic Maxwell equations. The scattering coefficient measures the exponential attenuation caused by scattering for every meter that a ray of light penetrates into a medium. Thus it is similar to the absorption coefficient, but the light is not transformed into other types of energy, it is just scattered away from the considered ray. Evidently some energy must also be scattered back into the considered ray. From similar considerations over the scattering coefficient Schuster constructed an equation describing the energy transfer in a scene.

Schuster originally considered an atmosphere modelled as a medium between two parallel planes. In a more general setting we have to model the directional tendency in the scattering. This means that we need more than just a scattering coefficient. Extending Schuster's theory, Schwarzschild [1906] introduced what we today call the *phase function*. This is a phenomenological function which describes the magnitude of scattering as a function of the angle with the direction of the considered ray of light. With this extension of the theory Schwarzschild formulated what we today refer to as the *radiative transfer equation* (sometimes abbreviated RTE). The full radiative transfer theory was introduced to graphics by Kajiya and Von Herzen [1984]. They suggested that the radiative transfer equation should be used for realistic rendering of volumes, and indeed the equation is today often referred to as the *volume rendering equation* in graphics. Since it is a phenomenological equation, that is, a mathematical model which was not originally derived from physics, we will first find out what it looks like. In its differential form the radiative transfer equation is [Chandrasekhar 1950]:

$$(\vec{\omega} \cdot \nabla) L(\boldsymbol{x}, \vec{\omega}) = -\sigma_t(\boldsymbol{x}) L(\boldsymbol{x}, \vec{\omega}) + \sigma_s(\boldsymbol{x}) \int_{4\pi} p(\boldsymbol{x}, \vec{\omega}', \vec{\omega}) L(\boldsymbol{x}, \vec{\omega}') \, \mathrm{d}\omega' + L_e(\boldsymbol{x}, \vec{\omega}) \ , \tag{6.1}$$

where $L(\boldsymbol{x}, \vec{\omega})$ is the *radiance* at $\boldsymbol{x}$ in the direction $\vec{\omega}$, the subscript e denotes emission, and σ_s, σ_a, and $\sigma_t = \sigma_s + \sigma_a$ are the scattering, absorption, and extinction coefficients respectively. The phase function p specifies the normalised distribution of the scattered light. Radiance is a radiometric quantity measured in energy flux per solid angle per projected area. The equation splits the directional derivative (left-hand side), that is, the change in radiance along a ray, into three terms (right-hand side): The first term denotes the exponential attenuation, the second denotes the in-scattering from all directions, and the third is an emission term.

To complete the theoretical understanding that we would like to have of the simplifications involved in the theory of light that we use, we will, in this chapter, relate the radiative transfer equation to the physical theories presented in the previous chapters.

It should be mentioned that there are many ways to generalise the radiative transfer equation. The same equation is used in neutron transport theory [Weinberg and Wigner 1958; Case and Zweifel 1967] where it is often called the *Boltzmann equation*. The Boltzmann equation also includes a term for variation of the radiance over time. Thus the radiative transfer equation (as defined by Equation 6.1) would be the stationary Boltzmann equation, and it can be generalised to include the same term for variation over time. Modified radiative transfer equations have also been proposed for radiative transfer in dense media [Goedecke 1977; Wen et al. 1990]. In the following we will only consider the conventional, stationary radiative transfer equation (6.1).

6.1 Scattering by a Particle

To introduce scattering in an electromagnetic field, we have to think a little like when we were constructing a Hamiltonian operator in quantum electrodynamics (cf. Chapter 3). We write the energy of a field containing scatterers as the sum of three terms: One describing the incident field $\boldsymbol{S}_i$, one describing the scattered field $\boldsymbol{S}_s$, and one describing the interaction between the two fields $\boldsymbol{S}_{\text{ext}}$.

As a consequence of Equation 5.12, another way to write the time-averaged Poynting vector is:

$$\boldsymbol{S}_{\text{avg}} = \frac{\varepsilon_0 c^2 \mu}{2} \text{Re}(\boldsymbol{E}_c \times \boldsymbol{H}_c^*) \ , \tag{6.2}$$

where ε_0 is the electric constant, c is the speed of light in vacuum, μ is the permeability, $\boldsymbol{E}_c$ and $\boldsymbol{H}_c$ are the time-harmonic electric and magnetic vectors, Re takes the real part of a complex quantity, and the asterisk * denotes the complex conjugate. The factor $\varepsilon_0 c^2 \mu$ cancels out if the material is not magnetic (cf. Equation 4.28).

Consider a particle in an electromagnetic field. To describe the effect, we split the field in two contributions: An incident field $(\boldsymbol{E}_i, \boldsymbol{H}_i)$ and a scattered field $(\boldsymbol{E}_s, \boldsymbol{H}_s)$. The sum of the two contributions makes up the total electromagnetic field:

$$\boldsymbol{E}_c = \boldsymbol{E}_i + \boldsymbol{E}_s \quad , \quad \boldsymbol{H}_c = \boldsymbol{H}_i + \boldsymbol{H}_s \ .$$

When we insert in Equation 6.2 to find the total time-averaged Poynting vector, we get the three terms just mentioned (incident, scattered, interaction):

$$\boldsymbol{S}_{\text{avg}} = \boldsymbol{S}_i + \boldsymbol{S}_s + \boldsymbol{S}_{\text{ext}} \ , \tag{6.3}$$

where

$$\boldsymbol{S}_i = \frac{\varepsilon_0 c^2 \mu}{2} \text{Re}(\boldsymbol{E}_i \times \boldsymbol{H}_i^*) \tag{6.4}$$

$$\boldsymbol{S}_s = \frac{\varepsilon_0 c^2 \mu}{2} \text{Re}(\boldsymbol{E}_s \times \boldsymbol{H}_s^*) \tag{6.5}$$

$$\boldsymbol{S}_{\text{ext}} = \frac{\varepsilon_0 c^2 \mu}{2} \text{Re}(\boldsymbol{E}_i \times \boldsymbol{H}_s^* + \boldsymbol{E}_s \times \boldsymbol{H}_i^*) \ . \tag{6.6}$$

To get an idea about the directions of the different vectors, we introduce a coordinate system. See Figure 6.1. The direction of the incident light defines an axis of incidence. Let us orient our coordinate system such that it has origin at the center of the particle and z-axis along the axis of incidence. The z-axis is then the *forward direction*. Together the forward direction and the direction in which we consider the scattered light defines the *scattering plane*. If the two

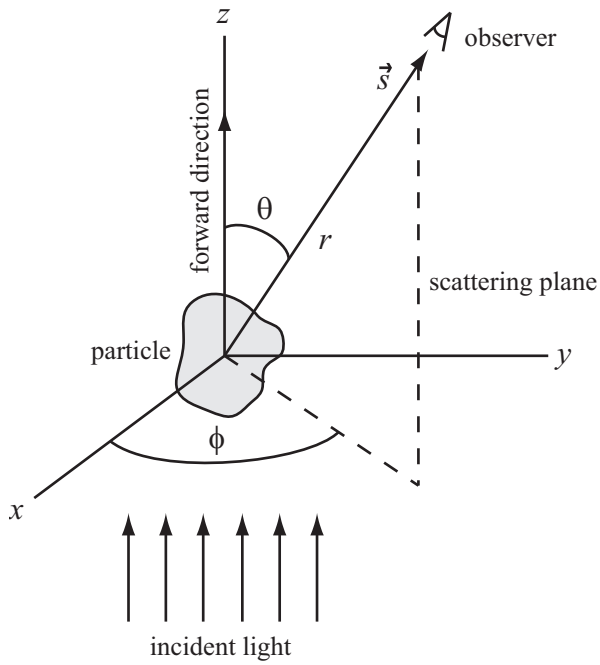


Figure 6.1: *Scattering by a particle.*

directions are parallel, we may choose any plane containing the axis of incidence. In this setting a spherical coordinate system (r, θ, ϕ) is oriented such that θ is the angle between the forward direction and the scattered direction, while ϕ is the angle between the x-axis and the scattering plane. This description of the setup follows that of Bohren and Huffman [1983, Sec. 3.2].

At a large distance from the particle (in the far field) the scattered field may be expressed as a spherical wave [Born and Wolf 1999; Bohren and Huffman 1983]:

$$\boldsymbol{E}_s(\boldsymbol{r}) = i\frac{e^{ikr}}{kr}\boldsymbol{Z}(\vec{s}) \quad , \quad \boldsymbol{H}_s(\boldsymbol{r}) = \frac{k}{\omega\mu} i\frac{e^{ikr}}{kr}\vec{s}\times\boldsymbol{Z}(\vec{s}) \ , \tag{6.7}$$

where r is the distance to the observer (which is assumed to be the same in all directions), ω is the angular frequency, k is the wave number, μ is the permeability, and both the index of refraction n and the permeability μ are those of the medium surrounding the particle. The vector $\boldsymbol{Z}$ denotes the strength of the light scattered in the direction $\vec{s}$. There must be a relation between the incident electric field $\boldsymbol{E}_i$ and the scattered field hidden in the directionally dependent *scattering vector function* $\boldsymbol{Z}(\vec{s})$. Let us try to formulate this relationship in the following paragraph.

As in reflection and refraction we consider two polarisations: One with the electric vector perpendicular to the scattering plane ($\perp$, TE) and one with the electric vector parallel to the scattering plane ($\parallel$, TM). Suppose we (in the case of the electric vector) denote the amplitudes of these two components by $E_{i\perp}$ and $E_{i\parallel}$ for the incident field, and by $E_{s\perp}$ and $E_{s\parallel}$ for the scattered field. Then we introduce a *scattering matrix* $\mathbf{S}(\theta, \phi)$ which in general depends on the angle θ between the forward direction and the direction of the scattered light as well as the angle ϕ which determines the location of the scattering plane. Putting all this notation together, we get a formula describing the relationship between the amplitudes of the scattered components and those of the incident components [van de Hulst 1957; Bohren and Huffman 1983]:

$$\begin{pmatrix} E_{s\parallel} \\ E_{s\perp} \end{pmatrix} = i\frac{e^{ik(r-z)}}{kr} \begin{pmatrix} S_2 & S_3 \\ S_4 & S_1 \end{pmatrix} \begin{pmatrix} E_{i\parallel} \\ E_{i\perp} \end{pmatrix} ,$$

where the four-component matrix is the scattering matrix $\mathbf{S}(\theta, \phi)$ and z is the distance from the light source to the origin of the coordinate system which was placed inside the particle. In the forward direction, when $\theta = 0$, the angle ϕ is of no consequence and we simply write $\mathbf{S}(0)$ for the scattering matrix. If the particle is a perfect sphere, we have $S_3 = S_4 = 0$ and in the forward direction also $S_1(0) = S_2(0) = S(0)$ [van de Hulst 1957].

The scattering matrix gives an expression for the scattering vector function $\boldsymbol{Z}$:

$$\boldsymbol{Z} = e^{-ikz} \begin{pmatrix} \vec{e}_{\parallel} & \vec{e}_{\perp} \end{pmatrix} \begin{pmatrix} S_2 & S_3 \\ S_4 & S_1 \end{pmatrix} \begin{pmatrix} E_{i\parallel} \\ E_{i\perp} \end{pmatrix} , \tag{6.8}$$

where $\vec{e}_{\parallel}$ and $\vec{e}_{\perp}$ are unit vectors parallel and perpendicular to the scattering plane respectively. They are chosen such that $\vec{e}_{\parallel} \times \vec{e}_{\perp} = \vec{s}$. Now that we have an expression connecting $\boldsymbol{Z}$ to the incident field, let us see how $\boldsymbol{Z}$ relates to the scattered energy. By insertion of the spherical wave equations (6.7) in the expression for the scattering Poynting vector (6.5), we get

$$\boldsymbol{S}_s = \frac{\varepsilon_0 c^2}{2\omega} \frac{e^{-2k''r}}{r^2} \mathrm{Re}\left(\frac{\boldsymbol{Z}(\vec{s}) \times (\boldsymbol{k} \times \boldsymbol{Z}(\vec{s}))^*}{kk^*} \right) . \tag{6.9}$$

For comparison the Poynting vector of the incident light is:

$$\boldsymbol{S}_i = \frac{\varepsilon_0 c^2}{2\omega} \mathrm{Re}(\boldsymbol{E}_{i0} \times (\boldsymbol{k} \times \boldsymbol{E}_{i0})^*) e^{-2k''z} .$$

In both these equations the wave vector is related to the direction of scattering by $\boldsymbol{k} = k\vec{s}$. Because the vectors are perpendicular (and using $k' \approx n'\omega/c$), we get the following magnitudes of the scattering and incident Poynting vectors:

$$\begin{aligned} |\boldsymbol{S}_s| &= \frac{\varepsilon_0 c}{2} n' \frac{e^{-2k''r}}{r^2} \frac{|\boldsymbol{Z}(\vec{s})|^2}{|k|^2} \\ |\boldsymbol{S}_i| &= \frac{\varepsilon_0 c}{2} n' |\boldsymbol{E}_{i0}|^2 e^{-2k''z} . \end{aligned}$$

Let us see if we can find out something about the ratio of scattered to incident light. We denote the surface of the particle A. The rate of energy transfer across a closed surface arbitrarily close to A is given by the integral [Bohren and Huffman 1983]

$$\int \boldsymbol{S}_{\text{avg}} \cdot \vec{n} \, \mathrm{d}A \ ,$$

where $\vec{n}$ denotes the outward surface normal. The total scattering by a particle is then an integral over the surface of the particle. We observe that there are two ways to consider a patch on the surface of the particle: one way is to consider an area ΔA, but another way is to describe the patch by a solid angle $\Delta\Omega$. We have

$$\Delta\Omega = \frac{\Delta A}{r^2} \ .$$

This means that the energy scattered by a particle is given by

$$W_s = \int \boldsymbol{S}_s \cdot \vec{s} \, \mathrm{d}A = \int_{4\pi} \frac{\varepsilon_0 c}{2} n' e^{-2k'' r} \frac{|\boldsymbol{Z}(\vec{s})|^2}{|k|^2} \, \mathrm{d}\Omega \ .$$

What we are after is the ratio of scattered to incident light. The conventional way of describing this ratio is to divide W_s by the incident energy flux $|\boldsymbol{S}_i|$. To make the two quantities look more alike, we use the weights of the two polarisations ($\parallel$ and $\perp$) instead of $E_{i\parallel}$ and $E_{i\perp}$ in the expression for $\boldsymbol{Z}$. Then it is possible to extract the magnitude of the incident field $|\boldsymbol{E}_{i0}|$ from $\boldsymbol{Z}$. Let us write $|\boldsymbol{E}_{i0}| e^{-ikz} \boldsymbol{X}(\vec{s}) = \boldsymbol{Z}(\vec{s})$. Then the ratio is

$$C_s = \frac{W_s}{|\boldsymbol{S}_i|} = e^{-2k'' r} \int_{4\pi} \frac{|\boldsymbol{X}(\vec{s})|^2}{|k|^2} \, \mathrm{d}\Omega \ . \tag{6.10}$$

A ratio of total scattered energy to energy incident per unit area gives an area as result. This particular area C_s is referred to as the *scattering cross section* of the particle. It is the area that would receive the same amount of energy as the particle scatters (to the distance r) if we subtend it normal to the incident light.

The exponential attenuation term $e^{-2k'' r}$ in the direction towards the observer is rather unfortunate. It means that the scattering cross section is not really an independent property of the particle. Nevertheless it is the generally accepted approach to use it as such. The approximation made in this context is that the absorption of the medium around the particle is small, thus we can neglect the exponential term or set r to the radius of a sphere bounding the particle. This assumption is not a good one, but since we want to use the scattering cross section for constructing macroscopic scattering properties, we will have to make do with it.

If we normalise the directionally dependant quantity under the integral, we have the phase function of the particle:

$$p = \frac{|\boldsymbol{X}^2|}{|k|^2 C_s} \quad .$$

This indicates that there is a connection to the radiative transfer equation (6.1). Before we get to the connection between the two worlds, we first need more than one particle.

6.2 Macroscopic Scattering

In the spirit of the previous section each particle that we introduce is modelled by a scattered electromagnetic field. Suppose we have N particles, then the total electromagnetic field is

$$\boldsymbol{E}_c = \boldsymbol{E}_i + \sum_{j=1}^{N} \boldsymbol{E}_{s,j} \quad , \quad \boldsymbol{H}_c = \boldsymbol{H}_i + \sum_{j=1}^{N} \boldsymbol{H}_{s,j} \quad .$$

When we insert this expression in the expression for the total time-averaged Poynting vector (6.2), we get the following three terms instead of Equations 6.4–6.6 [Bohren and Huffman 1983]:

$$\boldsymbol{S}_i = \frac{\varepsilon_0 c^2 \mu}{2} \mathrm{Re}(\boldsymbol{E}_i \times \boldsymbol{H}_i^*) \tag{6.11}$$

$$\boldsymbol{S}_s = \frac{\varepsilon_0 c^2 \mu}{2} \sum_{j=1}^{N} \sum_{l=1}^{N} \mathrm{Re}(\boldsymbol{E}_{s,j} \times \boldsymbol{H}_{s,l}^*) \tag{6.12}$$

$$\boldsymbol{S}_{\mathrm{ext}} = \frac{\varepsilon_0 c^2 \mu}{2} \sum_{j=1}^{N} \mathrm{Re}(\boldsymbol{E}_i \times \boldsymbol{H}_{s,j}^* + \boldsymbol{E}_s \times \boldsymbol{H}_{i,j}^*) \quad . \tag{6.13}$$

To generalise the results found for a single particle in the previous section, we could try to use a spherical wave to model the combined scattering effect of a collection of particles. Obviously the particles cannot all be placed at the origin, so we need something to handle the fact that they are placed at different positions in space.

Suppose we denote the wave vector of the incident light $\boldsymbol{k}_0$. In our single particle case we had $\boldsymbol{k}_0 \cdot \boldsymbol{x} = -kz$. Now we let $\boldsymbol{x}_j$ denote the position of particle j. Because we model the combined scattering as a spherical wave (far-field approximation), we are allowed to assume that the direction of the scattered

light $\vec{s}$ is the same for all particles. The wave vector of the scattered light is then $\boldsymbol{k}_1 = k\vec{s}$, and we let $\boldsymbol{K} = \boldsymbol{k}_1 - \boldsymbol{k}_0$ denote the phase change. In this terminology, which follows Champeney [1973], the scattering vector function $\boldsymbol{Z}$ is

$$\boldsymbol{Z}(\vec{s}) = |\boldsymbol{E}_{i0}| \sum_{j=1}^{N} e^{-i\boldsymbol{K}\cdot\boldsymbol{x}_j} \boldsymbol{X}_j(\vec{s}) \ . \tag{6.14}$$

Or, if we introduce a density function

$$\rho(\boldsymbol{x}, \vec{s}) = \sum_{j=1}^{N} \boldsymbol{X}_j(\vec{s}) \delta(\boldsymbol{x} - \boldsymbol{x}_j) \ ,$$

which involves point delta functions for the positions of the particles, we can write the scattering vector function as a Fourier transform

$$\boldsymbol{Z}(\vec{s}) = |\boldsymbol{E}_{i0}| \int \rho(\boldsymbol{x}, \vec{s}) e^{-i\boldsymbol{K}\cdot\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x} = |\boldsymbol{E}_{i0}| \boldsymbol{X}(\vec{s}) \ .$$

In this way we have packed all the difficulties into the function $\boldsymbol{X}(\vec{s})$ which now describes the total scattering by all the particles in the direction $\vec{s}$.

On immediate inspection the spherical wave approach seems to be a very limited point of view. As we have formulated it, it is only valid if we are considering a number of scatterers from far away. If we were standing in the midst of them, like in the middle of a cloud, the approximation would no longer be so good. This is not necessarily the case. If we replace $\rho(\boldsymbol{x}, \vec{s})$ by the number density $N(\boldsymbol{x})$ of scatterers per unit volume, and think of $\boldsymbol{X}(V, \vec{s})$ as the collected scattering of volume V in the direction $\boldsymbol{s}$, then the equations are still valid for a very small element of volume [Champeney 1973]. Thus when we let the volumes go to the limit of being points in a medium ($V \to \boldsymbol{x}$), the equations are also valid even if we are moving around inside the medium. If we are tracing a ray of light through a medium, it is comforting to know that we are able to express the scattering at a point $\boldsymbol{x}$ in the medium in the direction $\vec{s}$. Thus we have obtained our macroscopic scattering properties: They are the $\boldsymbol{X}(\boldsymbol{x}, \vec{s})$ vector function, or the phase function p before it is normalised if you prefer.

If we insist on a normalised phase function (and we often do), we need a macroscopic version of the scattering cross section. This is found by the same procedure as the one we used in the previous section to find the scattering cross section of a single particle. First we integrate the magnitude of the Poynting vector for the scattered field (6.12) over all solid angles. And then we divide by the magnitude of the Poynting vector for the incident field (6.11). In the far field the result is

$$C_{s,\mathrm{macro}} = \frac{W_s}{|\boldsymbol{S}_i|} = e^{-2k''(r-z)} \int_{4\pi} \left| \sum_{j=1}^{N} \sum_{l=1}^{N} \frac{\boldsymbol{X}_j(\vec{s}) \cdot \boldsymbol{X}_l^*(\vec{s})}{k^2} \right| \mathrm{d}\Omega \ .$$

This shows that we cannot simply add the scattering cross sections of the individual particles to get the correct macroscopic scattering cross section. Nevertheless, it is common practice to use this approximation. For comparison the approximation is

$$C_{s,\text{macro}} \approx e^{-2k''(r-z)} \sum_{j=1}^{N} \int_{4\pi} \frac{|\boldsymbol{X}_j(\vec{s})|^2}{|k|^2} \,\mathrm{d}\Omega \ . \tag{6.15}$$

It is difficult to say what the simplifying assumption is exactly. One way to explain it is to say that all secondary scattering events have been neglected. However, if we let the scattering cross section describe the scattering in an infinitesimal volume (as we did for the scattering function $\boldsymbol{X}$ before), we are still able to simulate multiple scattering in a macroscopic way. In this case we cannot simply say that all secondary scattering events have been neglected. They may be overestimated or underestimated due to a macroscopic scattering cross section which is slightly off target, but they are not neglected. Precisely how much the macroscopic scattering cross section is off target, when we neglect secondary scattering in the microscopic description, is not easy to say. It is also difficult to say if it is too large or too small. We may conjecture that it is probably too large if the particles are absorbing and too small if they are not. The general conception, which is probably true, is that the independent scattering approximation (6.15) is a good approximation as long as the particles are not too densely packed in the medium [van de Hulst 1957; Goedecke 1977].

The cross section concept is a little strange. Why represent the ratio of scattered to incident light in terms of an area? I am not really sure why this choice has been made in physics. Perhaps it is because it relates to the concept of number density. If C_s is the scattering cross section of a particle, then, in the independent scattering approximation, the macroscopic version of it would be C_s times the number of scattering cross sections per unit volume. Since the particles may have many sizes with different cross sections, the macroscopic scattering cross section should be an integral over particle sizes a. Let us denote it σ_s, we have

$$\sigma_s(\boldsymbol{x}) = \int_0^\infty C_s(\boldsymbol{x}, a) N(\boldsymbol{x}, a) \,\mathrm{d}a \ . \tag{6.16}$$

This quantity, σ_s, is Schuster's scattering coefficient as derived from Maxwell's equations. The number density in this expression is actually $N(\boldsymbol{x}, a)\,\mathrm{d}a$, while $N(\boldsymbol{x}, a)$ itself is a number density distribution. It is a distribution because it denotes the number of particles of size a in the range of particle sizes $\mathrm{d}a$. The scattering coefficient denotes the scattering per unit length that a ray suffers when penetrating the particulate medium. Thus the scattering coefficient is similar to the absorption coefficient described in Section 4.3 (cf. Equation 4.43). Another way to describe the macroscopic scattering properties of a medium is then by a scattering coefficient σ_s and a normalised phase function p.

It is convenient to introduce an *extinction coefficient* σ_t which is the sum of the scattering and absorption coefficients:

$$\sigma_t = \sigma_a + \sigma_s \ . \tag{6.17}$$

Having described all three of these coefficients as well as the normalised phase function, we now know where the optical properties used in the radiative transfer equation (6.1) come from. It remains to find out how the equation itself comes about. This is the subject of the following section.

6.3 The Radiative Transfer Equation

The radiometric quantity *radiance* is fundamental in radiative transfer. It is what the radiative transfer equation (6.1) is concerned with and it is defined (radiometrically) by

$$L = \frac{\mathrm{d}^3 Q}{\mathrm{d}t \, \mathrm{d}\omega \, \mathrm{d}A_\perp} = \frac{\mathrm{d}^2 \Phi}{\mathrm{d}\omega \, \mathrm{d}A \cos\theta} \ ,$$

where Q is energy, t is time, $\Phi = \mathrm{d}Q/\mathrm{d}t$ is energy flux, ω is solid angle, $A_\perp$ is projected area, A is area, and θ is the angle between the normal to the area $\mathrm{d}A$ and the considered direction. Radiance describes the flow of energy through a differential area $\mathrm{d}A$. The energy flows in a directional differential volume $\mathrm{d}\omega$ which is not necessarily normal to the area. The purpose of radiance is thus to describe the energy flow in a ray of light incident on a surface. The reason why we cannot simply assign an energy flux to a ray with no extension in space was considered in the historical remarks of Chapter 2. The reason is simply that energy needs a volume to flow in.

Since radiance models directional energy flow through an area, it is the perfect quantity for describing the energy flux moving through an image plane in the direction towards the eye. Thus radiance is what we compute for every pixel in an image plane, and it is what we use to find the colour that we want to assign to a pixel. As radiative transfer theory evolves around radiance, it is not surprising that the theory was introduced quite early in graphics.

To describe how the radiative transfer equation (6.1) comes about from Maxwell's equations, is a rather difficult subject. The difficulties arise because the equation was not based on a microscopic physical description from the beginning. It was created as a mathematical model based on intuitive arguments in order to describe the macroscopic phenomena that we experience when we observe the world (this is why we call it phenomenological). Nevertheless, the

radiative transfer equation has been incredibly successful in a wide range of applications. Therefore people have tried to derive it from Maxwell's equations many times [Ishimaru 1977], but it is a little like trying to fit a shoe in a glove.

The first and foremost problem is to define radiance in terms of electromagnetic quantities. The radiometric definition of radiance is heuristical. It is assumed to be of physical significance, but we have to take the integral of the radiance over all solid angles to obtain a quantity of unambiguous physical meaning [Wolf 1976; Fante 1981]. For this reason any definition of radiance which gives the correct result when integrated over all solid angles is an acceptable definition.

Since radiance is not uniquely defined, several different definitions have been proposed. Preisendorfer [1965, Chapter XIV] provided an early attempt to derive the radiative transfer theory from the electromagnetic field theory. However, it has been assumed from very early on that the radiative transfer quantity *fluence*, which is defined by

$$\phi(\boldsymbol{x}) = \int_{4\pi} L(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega \ , \tag{6.18}$$

is the same as the speed of light times the time-averaged electromagnetic energy density. That is, $\phi = cu_{\mathrm{avg}}$, where u_{avg} is the time-average of u as defined by Equation 4.6. This was assumed by Planck [1914] and many results in heat transfer rely on it. Unfortunately Preisendorfer's definition of radiance does not, in general, obey this basic assumption [Wolf 1976].

Another important quantity in radiative transfer theory is the *net flux* which is defined by

$$\boldsymbol{F}(\boldsymbol{x}) = \int_{4\pi} \vec{\omega} L(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega \ . \tag{6.19}$$

Wolf [1976] provided a different definition of radiance which satisfies both $\phi = cu_{\mathrm{avg}}$ and $\boldsymbol{F} = \boldsymbol{S}_{\mathrm{avg}}$. He showed that for a spatially homogeneous stationary electromagnetic field his definition, "with no approximation whatsoever" [Wolf 1976, p. 876], leads to the radiative transfer theory in free space. Based on this approach, Fante [1981] has provided a derivation of the radiative transfer equation for plane waves in a dielectric, and Sudarshan [1981] coupled the radiative transfer theory to quantum electrodynamics.

When the electromagnetic field is not in free space, the radiative transfer equation is only valid under certain conditions. The conditions depend on the definition of radiance that we choose, and on the equations that we would like the radiance quantity to fulfil. Fante [1981], for example, requires that radiance is traced in the direction normal to the surface of constant phase. This places additional constraints on the nature of the light and matter for which the radiative transfer equation is valid. If we take into account that energy does not

always flow in the direction normal to the surface of constant phase (using the theory proposed in the previous chapter), then we do not have to require that radiance follows the direction normal to the surface of constant phase. Thus it is difficult to say precisely how often the radiative transfer equation is valid (it depends on how you use it).

Yet another way to define radiance is in terms of a Fourier transform of the mutual coherence function [Ishimaru 1978]. Many authors have taken this approach, for example Ishimaru [1977; 1978] and Mishchenko [2002] (see also the references they provide). It is an approach which more clearly shows that the macroscopic optical properties in the radiative transfer equation are indeed the same as the macroscopic optical properties derived from the electromagnetic field theory. The most general derivation of the radiative transfer equation I am aware of is that of Mishchenko [2002; 2003] (the derivation is also available in the book by Mishchenko et al. [2006]). The conclusions of Mishchenko, as to the validity conditions of the radiative transfer equation, are consistent with the validity conditions that have previously been postulated (these are summarised, for example, by Goedecke [1977]). So it is highly probable that the validity conditions truly are the following:

- Particles and observer are all in the far-field zones of each other.
- Particles are statistically independent and randomly distributed throughout the medium.
- Particles scatter light independently.
- Particles in a small volume element scatter light at most once.

If we recall the simplifying assumptions involved in the independent scattering approximation (6.15), there is quite a good match between the validity conditions of the radiative transfer equation and those of the macroscopic optical properties. Thus we have a theory of macroscopic scattering which is fairly self-consistent.

Instead of taking part in the discussion about a definition of radiance in terms of electromagnetic quantities, I will stick with the indirect definition given by $\phi = cu_{\text{avg}}$ in Equation 6.18 and $F = \boldsymbol{S}_{\text{avg}}$ in Equation 6.19. The radiative transfer equation implies an interesting relationship between these two quantities which we will now derive.

Integrating the radiative transfer equation (6.1) over all solid angles $\mathrm{d}\omega$, the

following results since the phase function integrates to 1:

$$\int_{4\pi} (\vec{\omega} \cdot \nabla) L(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega = -(\sigma_t(\boldsymbol{x}) - \sigma_s(\boldsymbol{x})) \int_{4\pi} L(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega + Q(\boldsymbol{x}) \ . \tag{6.20}$$

Here Q is a source term defined by

$$Q(\boldsymbol{x}) = \int_{4\pi} L_e(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega \ .$$

Using the definitions of the fluence and net flux (6.18,6.19) as well as the relationship between the extinction, scattering, and absorption coefficients (6.17), simple considerations lead from this equation (6.20) to the relation

$$\nabla \cdot \boldsymbol{F}(\boldsymbol{x}) = -\sigma_a \phi(\boldsymbol{x}) + Q(\boldsymbol{x}) \ . \tag{6.21}$$

Using the basic assumptions $\phi = c u_{\mathrm{avg}}$ and $\boldsymbol{F} = \boldsymbol{S}_{\mathrm{avg}}$, we arrive at at the following relationship:

$$\nabla \cdot \boldsymbol{S}_{\mathrm{avg}}(\boldsymbol{x}) = -\sigma_a c u_{\mathrm{avg}}(\boldsymbol{x}) + Q(\boldsymbol{x}) \ . \tag{6.22}$$

This equation is a macroscopic time-averaged version of the differential equation for the conservation of electromagnetic energy (4.7). Intuitively, a comparison of the two equations (4.7,6.22) suggests that if our macroscopic optical properties (σ_a and L_e) faithfully model the microscopic properties, Equation 6.22 probably holds in most cases. Hopefully this gives some confidence in the validity of the radiative transfer equation. In the next section we will describe how the radiative transfer equation is evaluated in practice.

6.4 Rendering Volumes

In Chapter 5 we found a way to trace energy through matter. This was all we needed to render materials with a continuous interior. However, when a material is composed of millions of microscopic particles, it is no longer feasible to model the surface of every particle. Therefore we have, in this chapter, introduced macroscopic scattering, and in the previous section we saw that the radiative transfer equation (6.1) is a fairly self-consistent way to describe macroscopic scattering given an appropriate set of macroscopic scattering properties for a material.

The radiative transfer equation describes the relation between incident, scattered, and absorbed radiance at a point in a medium. Since radiance is a quantity based on energy, we assume that it follows the flow of energy through

a medium. Thus we trace rays using the eikonal equation (5.7) as described in Chapter 5. However, we have to be careful and take into account that radiance denotes flux per solid angle per projected area. When we follow a ray of light through a material, we have to take into account that the radiance may change along the ray. Let us see how radiance changes upon refraction.

Consider a ray of light in a medium with refractive index n_1 incident on a surface patch $\mathrm{d}A$ of a medium with refractive index n_2. Due to energy conservation at the boundary, the following condition must hold:

$$L_i \cos\theta_i \,\mathrm{d}A \,\mathrm{d}\omega_i = L_r \cos\theta_r \,\mathrm{d}A \,\mathrm{d}\omega_r + L_t \cos\theta_t \,\mathrm{d}A \,\mathrm{d}\omega_t \ ,$$

where L_i, L_r, and L_t are the incident, reflected, and transmitted radiances and likewise θ_i, θ_r, and θ_t are the angle of incidence, the angle of reflection, and the angle of refraction. In spherical coordinates the solid angles are defined by

$$\begin{aligned} \mathrm{d}\omega_i &= \sin\theta_i \,\mathrm{d}\theta_i \,\mathrm{d}\phi_i \\ \mathrm{d}\omega_r &= \sin\theta_r \,\mathrm{d}\theta_r \,\mathrm{d}\phi_r \\ \mathrm{d}\omega_t &= \sin\theta_t \,\mathrm{d}\theta_t \,\mathrm{d}\phi_t \ . \end{aligned}$$

Using the law of reflection $\theta_i = \theta_r$ and the fact that both the reflected and transmitted rays lie in the plane of incidence $\phi_i = \phi_r = \phi_t$ (cf. Section 4.4), the boundary condition becomes

$$L_i \cos\theta_i \sin\theta_i \,\mathrm{d}\theta_i = L_r \cos\theta_i \sin\theta_i \,\mathrm{d}\theta_i + L_t \cos\theta_t \sin\theta_t \,\mathrm{d}\theta_t \ .$$

To find the transmitted angle, we have to consider the direction of the transmitted ray.

If we use the ray tracing scheme described in the previous chapter, we will split the ray in two directions: One for the TE component and one for the TM component. Using the directions given by Equations 5.16 and 5.17, we find

$$\begin{aligned} \sin\theta_t^{\perp} &= \frac{n_1'}{n_2'} \sin\theta_i \\ \cos\theta_t^{\perp} \,\mathrm{d}\theta_t^{\perp} &= \frac{n_1'}{n_2'} \cos\theta_i \,\mathrm{d}\theta_i \\ \sin\theta_t^{\parallel} &= \frac{(n_2'^2 - n_2''^2)n_1' + 2n_2'n_2''n_1''}{n_2'|n_2|^2} \sin\theta_i \\ \cos\theta_t^{\parallel} \,\mathrm{d}\theta_t^{\parallel} &= \frac{(n_2'^2 - n_2''^2)n_1' + 2n_2'n_2''n_1''}{n_2'|n_2|^2} \cos\theta_i \,\mathrm{d}\theta_i \ . \end{aligned}$$

With this result the boundary condition is

$$L_i = RL_i + T_{\perp} \left(\frac{n_1'}{n_2'}\right)^2 L_i + T_{\parallel} \left(\frac{(n_2'^2 - n_2''^2)n_1' + 2n_2'n_2''n_1''}{n_2'|n_2|^2}\right)^2 L_i \ , \qquad (6.23)$$

where we have used the Fresnel reflectance and transmittance described in Section 4.4. If both media are non-absorbing (or weak absorbers), the equation simplifies to

$$L_i = RL_i + T\left(\frac{n_1'}{n_2'}\right)^2 L_i \ . \tag{6.24}$$

These equations show that radiance is not constant along a ray of light. If we move along a ray through a heterogeneous medium, the radiance should be modified. Generalizing the result above to the interior of a medium, we approximately have [Preisendorfer 1965; Ishimaru 1978]

$$\frac{L_1}{n_1'^2}\mathrm{d}A_1 = \frac{L_2}{n_2'^2}\mathrm{d}A_2 \ ,$$

where the subscripts 1 and 2 denote two locations along a ray of light. This means that the quantity $L_1/n_1'^2$ is (approximately) constant along the ray. If we store $L_1/n_1'^2$ before tracing a ray from one point in a medium, then the radiance at the destination point is $L_2 = n_2'^2 L_1/n_1'^2$. Now that we know how radiance behaves as we move along a ray, we are ready to evaluate the radiative transfer equation (6.1) using ray tracing.

Rendering realistic images using the radiative transfer equation, was first proposed by Kajiya and Von Herzen [1984]. Most rendering algorithms uses approximate evaluation schemes to gain speed, we will look at approximate methods in Chapter 7. Before we start making approximations, we should learn a general way of evaluating the radiative transfer equation. The remainder of this section is a short account of *Monte Carlo path tracing*, which is a sampling-based rendering algorithm that works in general. More information is available in the book by Pharr and Humphreys [2004]. Unfortunately Pharr and Humphreys stop their treatment of volume rendering at single scattering. Evaluation of the general case has been described by Pattanaik and Mudur [1993].

The emission term in the radiative transfer equation is just an added constant. It is not difficult to include it, but it makes the equations rather long. Therefore we leave out the emission term in the following. The general approach is as follows. We first parameterise the radiative transfer equation using the distance s' that we have moved along a path into a medium. We have (for a non-emitter):

$$\frac{\mathrm{d}L(s')}{\mathrm{d}s'} + \sigma_t(s')L(s') = \sigma_s(s')\int_{4\pi} p(s', \vec{\omega}', \vec{\omega})L(s', \vec{\omega}')\,\mathrm{d}\omega' \ , \tag{6.25}$$

where $\vec{\omega}$ denotes the tangential direction of the path at the distance s' along the path.

The parameterised equation (6.25) is a linear, first-order, ordinary differential equation (where $\sigma_t(s')$ is a variable coefficient). One way to solve such an

equation is by means of an integration factor:

$$T_r(s', s) = \exp\left(-\int_{s'}^{s} \sigma_t(t)\,\mathrm{d}t\right) \quad . \tag{6.26}$$

With this factor the equation (6.25) translates to

$$\frac{\mathrm{d}}{\mathrm{d}s'}\left(T_r(s', s)L(s')\right) = T_r(s', s)\sigma_s(s') \int_{4\pi} p(s', \vec{\omega}', \vec{\omega})L(s', \vec{\omega}')\,\mathrm{d}\omega' \quad . \tag{6.27}$$

Note that $T_r(s, s) = 1$. Then by integration along the ray from the surface $s' = 0$ to the considered location in the medium $s' = s$, the equation (6.27) attains the form:

$$L(s) = T_r(0, s)L(0) + \int_0^s T_r(s', s)\sigma_s(s') \int_{4\pi} p(s', \vec{\omega}', \vec{\omega})L(s', \vec{\omega}')\,\mathrm{d}\omega'\,\mathrm{d}s' \quad . \tag{6.28}$$

For convenience some of the different mathematical quantities encountered in this derivation have been given appropriate names in radiative transfer theory [Chandrasekhar 1950]. The distance s (or s') traveled along the ray inside the medium, is referred to as the *depth*. The integral in Equation 6.26 is called the *optical thickness* and is denoted by the symbol τ, that is,

$$\tau(s', s) = \int_{s'}^{s} \sigma_t(t)\,\mathrm{d}t \quad .$$

The integration factor itself $T_r(s', s) = e^{-\tau(s', s)}$ is sometimes referred to as the *beam* (or path) *transmittance*. Finally, the first term on the right hand side of the formal solution (6.28) for the radiative transfer equation (6.1) is referred to as the *direct transmission term* whereas the second term is called the *diffusion term*. Realistic rendering is all about evaluating these terms in various kinds of ways.

To sample the equation (6.28) correctly using Monte Carlo path tracing, we take a look at the direct transmission term $T_r(0, s)L(0)$. For this term we need to evaluate the optical thickness $\tau(0, s)$. If the medium is homogeneous, this optical thickness is simply given by

$$\tau(0, s) = \sigma_t s \quad . \tag{6.29}$$

For heterogeneous media, we use Monte Carlo sampling. The optical thickness is found quite efficiently by the estimator:

$$\sum_{i=0}^{N-1} \sigma_t(t_i)\Delta t \quad , \tag{6.30}$$

where Δt is the step size (given as user input) and t_i are locations along the ray found using a single random variable $\xi \in [0, 1[$:

$$t_i = \frac{\xi + i}{N} s \ . \tag{6.31}$$

The number N of locations along the ray is found using the depth s which is the distance to the next surface, that is, $N = \lfloor s/\Delta t \rfloor$. Having estimated the optical thickness $\tau(0, s)$, the beam transmittance $T_r(0, s)$ is easily found and multiplied by the amount of radiance $L(0)$ to reveal the direct transmission term. The radiance $L(0)$ is the radiance which contributes to the ray at the surface of the medium.

Evaluating the diffusion term is more involved. What we need is an estimator of the form

$$\frac{1}{N} \sum_{j=0}^{N-1} \frac{T_r(s'_j, s)\sigma_s(s'_j)J(s'_j)}{\mathrm{pdf}(s'_j)} \ , \tag{6.32}$$

where the probability distribution function (pdf) preferably cancels out the transmittance. The source function

$$J(s') = \int_{4\pi} p(s', \vec{\omega}', \vec{\omega}) L(s', \vec{\omega}') \, \mathrm{d}\omega' \tag{6.33}$$

is evaluated using a distribution of samples over the entire unit sphere.

To sample the probability distribution function in Equation 6.32:

$$\mathrm{pdf}(s'_j) = \sigma_t(s'_j) T_r(s'_j, s) \ ,$$

we use the cumulative probability of an interaction along the ray. An interaction is either scattering according to the source function (6.33) or absorption. According to the cumulative probability of an interaction, an interaction occurs when [Pattanaik and Mudur 1993]

$$\ln(\xi_j) + \tau(s'_j, s) = 0 \ ,$$

where $\xi_j \in [0, 1[$ is a random variable for sample j. The depth of the sample is easily found for homogeneous media, where we have $\tau(s'_j, s) = (s - s'_j)\sigma_t$, which gives

$$s'_j = s + \frac{\ln(\xi_j)}{\sigma_t} \ . \tag{6.34}$$

If $s'_j < 0$, there is no interaction for sample j. For heterogeneous media we have to step along the ray to find out where the optical thickness matches the event. Starting at t_1 in $t_i = s - i\Delta t$, we step along the ray by incrementing i, and

if $\ln(\xi_j) + \tau(t_i, s) > 0$, we stop and compute the location of the interaction as follows [Pattanaik and Mudur 1993]:

$$s'_j = (i-1)\Delta t - \frac{\tau(t_{i-1}, s) + \ln(\xi_j)}{\sigma_t(t_i)} \quad . \tag{6.35}$$

Here it is assumed that $\sigma_t(t_i)$ is approximately constant in steps of size Δt along the ray, such that $\sigma_t(s'_j) \approx \sigma_t(t_i)$. If we end up with $t_i \leq 0$ before the other criteria is fulfilled, there is no interaction for sample j. The optical thicknesses needed in order to find s'_j are evaluated in the same way as when we evaluated $\tau(0, s)$ only with $s - t_i$ in Equation 6.31 instead of s.

Finally, the scattering coefficient $\sigma_s(s'_j)$ in the estimator (6.32) and the extinction coefficient $\sigma_t(s'_j)$ in the probability distribution function (pdf) are canceled out by means of a Russian roulette. At every interaction a Russian roulette is carried out using the *scattering albedo*, which is defined by

$$\alpha(s'_j) = \sigma_s(s'_j)/\sigma_t(s'_j) \quad ,$$

as the probability of a scattering event.

Using the sampling scheme described above, the estimator (6.32) becomes:

$$\frac{1}{N}\sum_{j=0}^{N-1} J(s'_j) = \frac{1}{NM}\sum_{j=0}^{N-1}\sum_{k=0}^{M-1} \frac{p(s'_j, \vec{\omega}'_k, \vec{\omega}) L(s'_j, \vec{\omega}'_k)}{\mathrm{pdf}(\vec{\omega}'_k)} \tag{6.36}$$

for $\xi < \alpha(s'_j)$ and 0 otherwise. For an isotropic phase function, sampling a uniform distribution over the unit sphere leaves only $L(s'_j, \vec{\omega}'_k)$ in the sum.

To summarise the algorithm, a ray is traced from an observer through a scene, when it refracts into a participating medium (that is, a scattering material), we do the following:

1. If the medium is a strong absorber, we trace two refracted rays (one for the TE component and one for the TM component). Otherwise we trace just one refracted ray. The radiance carried along the refracted ray is corrected according to the boundary condition (Equation 6.23 or Equation 6.24).
2. The tracing of the refracted ray gives the depth s to the next surface.
3. The radiance $L(0)$ which contributes to the ray at the surface is found by tracing new rays in the directions of reflection and refraction. If $L(0)$ is not too small, we evaluate the direct transmission term:
   (a) The optical thickness $\tau(0, s)$ is estimated using Equations 6.31 and 6.30 (or Equation 6.29 for homogeneous media).

(b) The direct transmission term $T_r(0,s)L(0)$ is found using the optical thickness $\tau(0,s)$ (see Equation 6.26) and the radiance which contributes to the ray at the surface $L(0)$.

4. For every diffusion term sample $j = 1,\ldots,N$, a sample depth s'_j is found using Equation 6.35 (or Equation 6.34 for homogeneous media).

5. For the samples $s'_j > 0$, a Russian roulette is done using the scattering albedo $\alpha(s'_j)$. For $\xi < \alpha(s'_j)$, where $\xi \in [0,1[$ is a random variable, there is a scattering event.

6. For every scattering event, the phase function $p(s'_j, \vec{\omega}'_k, \vec{\omega})$ is evaluated in M sampled directions $\vec{\omega}'_k$ with $k = 1,\ldots,M$. Likewise M new rays are traced at the position s'_j in the directions $\vec{\omega}'_k$ to obtain the radiances $L(s'_j, \vec{\omega}_k)$. Using Equation 6.36 this gives an estimate of the diffusion term.

7. Finally, the direct transmission term and the diffusion term are added to get the radiance emergent at the surface $L(s)$.

The numbers of samples chosen are often $N = 1$ and $M = 1$. Then we get a very noisy sample image rather quickly. Another sample is then rendered and this is averaged with the previous one. The next sample is weighted by one third and added to the other two samples which are weighted by two thirds and so on. In this way the image will improve itself over time. Even so, the Monte Carlo path tracing procedure spawns a formidable number of rays. It is very slow, but it is nice to use it for computing reference images. The images in Part IV are rendered using this approach with only few modifications to get a speed-up. In the next chapter we will investigate rendering techniques which are more approximate and therefore also faster.

Chapter 7

Surface and Diffusion Models

take the obvious things of every-day life, you will find them wonderfully complex as soon as you begin to go beneath the surface

D. Avery, from *Cultural Value of Science*

Mainstream photo-realistic rendering is based on radiative transfer theory. In Section 6.4 we saw that it is not so difficult to describe a general algorithm for realistic rendering. The real difficulties emerge when we want to speed up the computation. Since the radiative transfer equation (6.1), in general, is expensive to evaluate, many rendering algorithms do not evaluate it directly.

Frequently we are only interested in the radiance emergent on the surface of the objects in a scene. Therefore a phenomenological equation which only concerns radiance emergent on surfaces is commonly employed as a less expensive alternative to the volume rendering described by the radiative transfer equation. This equation is referred to as the *rendering equation* and it was introduced to graphics by Kajiya [1986].

One of the goals of this thesis is to find the physical origins of the material properties that we use and measure in graphics. In the previous chapter we saw how the macroscopic optical properties relate the radiative transfer equation to the physical theories of light. Unfortunately the rendering equation does not use the same macroscopic optical properties. To come closer to the goal, we will investigate, in this chapter, how the rendering equation relates to the radiative transfer equation.

The nomenclature, which we use in graphics for the phenomenological theory of radiative transfer between surfaces, originates in the field of optical radiation measurement with the work of Nicodemus et al. [1977]. The rendering equation is

$$L(\boldsymbol{x}_o, \vec{\omega}_o) = \int_A \int_{2\pi} S(\boldsymbol{x}_i, \vec{\omega}_i; \boldsymbol{x}_o, \vec{\omega}_o) L(\boldsymbol{x}_i, \vec{\omega}_i) \cos\theta \, \mathrm{d}\omega_i \, \mathrm{d}A + L_e(\boldsymbol{x}_o, \vec{\omega}_o) \ . \quad (7.1)$$

where $L(\boldsymbol{x}_o, \vec{\omega}_o)$ is the outgoing (or emergent) radiance in the direction $\vec{\omega}_o$ from the location x_o on the surface A of a medium, $L(\boldsymbol{x}_i, \vec{\omega}_i)$ is the radiance incident on the surface A at the location $\boldsymbol{x}_i$ from the direction $\vec{\omega}_i$, and $\cos\theta = \vec{\omega}_i \cdot \vec{n}$ is the angle between the surface normal $\vec{n}$ and the direction $\vec{\omega}_i$ toward the incident light. The function S is called a *Bidirectional Scattering-Surface Reflectance-Distribution Function* (BSSRDF) and L_e is an emission term.

Essentially the BSSRDF describes the fraction of light that a ray entering the medium at the surface location $\boldsymbol{x}_i$ from the direction $\vec{\omega}_i$ will contribute to the ray emerging at the surface location $\boldsymbol{x}_o$ in the direction $\vec{\omega}_o$. This is an immensely complicated function. Consider an element of flux $\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)$ incident on a surface location $\boldsymbol{x}_i$ from the direction $\vec{\omega}_i$, within the element of solid angle $\mathrm{d}\omega_i$. Let $\mathrm{d}L_r(\boldsymbol{x}_o, \vec{\omega}_o)$ denote the element of emergent radiance at the surface location $\boldsymbol{x}_o$ in the direction $\vec{\omega}_o$ which is due to the incident flux $\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)$. The assumption underlying the rendering equation (7.1) is that, for all incident and outgoing directions and locations, the emergent radiance $\mathrm{d}L_r(\boldsymbol{x}_o, \vec{\omega}_o)$ is proportional to the incident flux $\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)$. That is, [Nicodemus et al. 1977]

$$\frac{\mathrm{d}L_r(\boldsymbol{x}_o, \vec{\omega}_o)}{\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)} = S(\boldsymbol{x}_i, \vec{\omega}_i; \boldsymbol{x}_o, \vec{\omega}_o) \ . \quad (7.2)$$

The subscript r denotes reflectance. It has been added because the equation does not include the emission term.

Preisendorfer [1965] has shown that the rendering equation (7.1) for non-emitters ($L_e = 0$) follows from the radiative transfer equation (6.1). Of course, Preisendorfer does not call it "the rendering equation", he calls it "the global version of the continuous formulation of radiative transfer theory", but it is the same equation. The assumptions he uses are that the flux is incident and emergent from an arbitrary convex continuous medium with a constant index of refraction, and illuminated by a steady radiance distribution of arbitrary directional structure at each point of its boundary [Preisendorfer 1965, Sec. 22]. It is not assumed that the scattering properties are constant (thus the medium may be heterogeneous with respect to the scattering properties, but the rays of light follow straight line paths). Later he shows that the underlying assumption (7.2) is also true for flux incident and emergent on a connected concave continuous medium [Preisendorfer 1965, Sec. 25].

To derive the rendering equation from the radiative transfer equation Preisndorfer [1965] uses an operator $\boldsymbol{S}^j$ with $j = 1, 2, \ldots$. The operator denotes scattering events inside the medium. If we denote the direct transmission term $L^0 = T_r(0, s)L(0)$, the operator works such that $L^1 = \boldsymbol{S}^1 L^0$ is the emergent radiance if we include the light that has been scattered once before entering the ray. The assertion used to complete the proof is that the total emergent radiance L is

$$L = L^0 + \sum_{j=1}^{\infty} \boldsymbol{S}^j L^0 \ .$$

This is most probably true in general.

Finally, an interesting thing to note is the connection that Preisendorfer [1965] finds between the phase function and the BSSRDF. Consider flux incident on a medium from the direction $\vec{\omega}_i$, and emergent in the directions $\vec{\omega}_o$. Letting the medium shrink to a point, Preisendorfer finds that the BSSRDF, in the limit, is the same as $\sigma_s p(\vec{\omega}_i, \vec{\omega}_o)$, where σ_s is the scattering coefficient and p is the phase function. Thus there is certainly a relationship between the macroscopic optical properties used in volume rendering and those used in surface rendering.

The radiative transfer equation is the same as the equation used in stationary neutron transport theory. Therefore many results concerning neutron diffusion are also applicable in radiative transfer theory. The following two sections comprise a short account of diffusion theory, and one of the popular rendering technique called *subsurface scattering*. In subsurface scattering a BSSRDF is often derived using diffusion theory. Subsurface scattering is interesting in the context of this thesis because it shows how the macroscopic optical properties are directly involved when we derive BSSRDFs theoretically. Fick's law of diffusion is a key element in the derivation of the dipole approximation used in subsurface scattering. Let us, therefore, first investigate under which simplifying assumptions Fick's law of diffusion is valid.

7.1 Fick's Law of Diffusion

The purpose of diffusion theory is to provide an efficient approximative alternative to Monte Carlo evaluation of the radiative transfer equation (6.1). To make the problem simpler, we only consider homogeneous, non-emitting media, and we make the very common assumption that scattering is rotationally invariant. This means that

$$p(\boldsymbol{x}, \vec{\omega}', \vec{\omega}) = p(\vec{\omega}' \cdot \vec{\omega}) = p(\cos\theta_0) \ .$$

Due to the differentiation theorem for Fourier transforms, the directional derivative in the radiative transfer equation (6.1) is much easier to handle if we take the three-dimensional Fourier transform. We let $\tilde{L}$ denote the Fourier transformed radiance, such that

$$L(\boldsymbol{x}, \vec{\omega}) = \frac{1}{(2\pi)^3} \int \tilde{L}(\boldsymbol{u}, \vec{\omega}) e^{i\boldsymbol{u}\cdot\boldsymbol{x}} \, \mathrm{d}\boldsymbol{u} \ .$$

Inserting this expression for L in the radiative transfer equation (6.1), we get the equation in terms of the Fourier transformed radiance (for a non-emitter):

$$(\sigma_t + i\boldsymbol{u} \cdot \vec{\omega})\tilde{L}(\boldsymbol{u}, \vec{\omega}) = \sigma_s \int_{4\pi} p(\cos\theta_0)\tilde{L}(\boldsymbol{u}, \vec{\omega}') \, \mathrm{d}\omega' \ , \tag{7.3}$$

This equation reveals that the Fourier transformed radiance $\tilde{L}$ may be taken as a function of $u = |\boldsymbol{u}|$ and $\cos\theta = (\boldsymbol{u} \cdot \vec{\omega})/u$ [Waller 1946].

Because we assumed rotationally invariant scattering, we can expand the phase function in terms of a series of Legendre polynomials:

$$p(\cos\theta_0) = \sum_{n=0}^{\infty} \frac{2n+1}{4\pi} p_n P_n(\cos\theta_0) \ , \tag{7.4}$$

where the expansion coefficients p_n are determined by the orthogonality relations for the Legendre polynomials P_n. This means that

$$p_n = \int_{4\pi} p(\cos\theta_0) P_n(\cos\theta_0) \, \mathrm{d}\omega \ .$$

The first two values for the Legendre polynomials are $P_0(\mu) = 1$ and $P_1(\mu) = \mu$. This means that we can interpret the first two expansion coefficients as follows:

$$\begin{array}{rrl} p_0 = & \int_{4\pi} p(\cos\theta_0) \, \mathrm{d}\omega & = 1 \\ p_1 = & \int_{4\pi} p(\cos\theta_0) \cos\theta_0 \, \mathrm{d}\omega & = g \ , \end{array} \tag{7.5}$$

where g is called the *asymmetry parameter* and the rightmost equality in the first equation follows because the phase function is normalised. The asymmetry parameter denotes the weighted mean cosine of the scattering angle with the phase function as weighting function. In a sense it describes the shape of the phase function: If $g = 1$, all light is scattered in the forward direction; if $g = -1$, all light is backscattered; and if $g = 0$, the phase function is perfectly isotropic.

To explain the angles referred to in the following, we note that the elements of solid angles $\mathrm{d}\omega$ and $\mathrm{d}\omega'$ have a representation in spherical coordinates such that

$$\begin{aligned} \mathrm{d}\omega &= \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\varphi \\ \mathrm{d}\omega' &= \sin\theta' \, \mathrm{d}\theta' \, \mathrm{d}\varphi' \ . \end{aligned}$$

The addition theorem for Legendre polynomials says that

$$P_n(\cos\theta_0) = P_n(\cos\theta)P_n(\cos\theta') + 2\sum_{m=-n}^{n}\frac{(n-m)!}{(n+m)!}P_n^m(\cos\theta)P_n^m(\cos\theta')\cos(m(\varphi-\varphi')) \ , \qquad (7.6)$$

where P_n^m are the associated Legendre polynomials. Using the addition theorem, the expansion of the phase function (7.4) can be expressed as a function of $\cos\theta$ and $\cos\theta'$. The second term in the addition theorem (7.6) cancels out if we integrate over φ from 0 to 2π since, as m is an integer, we have

$$\int_0^{2\pi}\cos(m(\varphi-\varphi'))\,\mathrm{d}\varphi = 0 \ .$$

Thus if we insert the expansion of the phase function (7.4) in Equation 7.3, use the addition theorem (7.6), and integrate both sides of the equation over φ from 0 to 2π, we obtain

$$(\sigma_t + i\boldsymbol{u}\cdot\vec{\omega})2\pi\tilde{L}(\boldsymbol{u},\vec{\omega}) = \sigma_s\int_{4\pi}\sum_{n=0}^{\infty}\frac{2n+1}{4\pi}p_nP_n(\cos\theta)P_n(\cos\theta')2\pi\tilde{L}(\boldsymbol{u},\vec{\omega}')\,\mathrm{d}\omega' \ . \qquad (7.7)$$

As mentioned before, $\tilde{L}$ may be taken as a function of u and $\cos\theta$. Therefore we choose

$$\tilde{L}(u,\cos\theta) = 2\pi\tilde{L}(\boldsymbol{u},\vec{\omega}) \ .$$

Using this relation, and replacing the integral over all solid angles with an integral over spherical coordinates, Equation 7.7 takes the following form:

$$(\sigma_t + iu\cos\theta)\tilde{L}(u,\cos\theta) = \sigma_s\sum_{n=0}^{\infty}\frac{2n+1}{2}p_nP_n(\cos\theta) \times\int_{-1}^{1}P_n(\cos\theta')\tilde{L}(u,\cos\theta')\,\mathrm{d}(\cos\theta') \ . \qquad (7.8)$$

The orthogonality relations for Legendre polynomials says that

$$\int_{-1}^{1}P_n(\mu)P_m(\mu)\,\mathrm{d}\mu = \begin{cases} 2/(2n+1) & \text{for } n=m \\ 0 & \text{otherwise} \end{cases} \ . \qquad (7.9)$$

Multiplying Equation 7.8 by $P_1(\cos\theta) = \cos\theta$, and using the orthogonality relations (7.9), we obtain the following after integration over $\cos\theta$ from -1 to 1 at both sides of the equation:

$$\int_{-1}^{1}\cos\theta(\sigma_t + iu\cos\theta)\tilde{L}(u,\cos\theta)\,\mathrm{d}(\cos\theta) = \sigma_s p_1\int_{-1}^{1}P_1(\cos\theta')\tilde{L}(u,\cos\theta')\,\mathrm{d}(\cos\theta') \ . \qquad (7.10)$$

In the same way that we expanded the phase function in terms of a series of Legendre polynomials we can also expand the Fourier transformed radiance $\tilde{L}$ in Legendre polynomials. We have

$$\tilde{L}(u,\mu) = \sum_{n=0}^{\infty} \tilde{L}_n(u) P_n(\mu) \ ,$$

where $\tilde{L}_n$ are the expansion coefficients. Using this expansion with both $\mu = \cos\theta$ and $\mu = \cos\theta'$, and the orthogonality relations once again (7.9), Equation 7.10 becomes

$$\tilde{L}_1(u) = -\frac{1}{\sigma_t - g\sigma_s} \int_{-1}^{1} iu \cos^2\theta \, \tilde{L}(u, \cos\theta) \, \mathrm{d}(\cos\theta) \ , \tag{7.11}$$

where $\tilde{L}_1$ is the second coefficient in the Legendre polynomials expansion of $\tilde{L}$, and the phase function expansion coefficient p_1 has been replaced by the asymmetry parameter g (7.5).

The mean square cosine of the Fourier transformed radiance distribution is defined by

$$\langle \cos^2\theta \rangle_{\mathrm{avg}} = \frac{\int_{4\pi} \cos^2\theta \ \tilde{L}(u, \cos\theta) \, \mathrm{d}\omega}{\int_{4\pi} \tilde{L}(u, \cos\theta) \, \mathrm{d}\omega} \ . \tag{7.12}$$

Here it should be noted that if (and only if) $\langle \cos^2\theta \rangle_{\mathrm{avg}}$ is independent of the position in $\boldsymbol{u}$-space, we may legally replace $\tilde{L}(u, \cos\theta)$ with $L(\boldsymbol{x}, \vec{\omega})$ in this definition. Using this mean square cosine quantity with Equation 7.11, we obtain

$$\tilde{L}_1(u) = -\frac{iu\langle \cos^2\theta \rangle_{\mathrm{avg}}}{\sigma_t - g\sigma_s} \tilde{L}_0(u) \ , \tag{7.13}$$

where $\tilde{L}_0$ is the first coefficient in the Legendre polynomials expansion of $\tilde{L}$.

If we, in turn, place $\boldsymbol{u}$ along each axis in $\boldsymbol{u}$-space and replace $\cos\theta$ by the direction cosines:

$$(\cos\alpha, \cos\beta, \cos\gamma) = \vec{\omega} \ , \tag{7.14}$$

the result is the following three-component vector equation

$$\begin{pmatrix} \tilde{L}_1(u_x) \\ \tilde{L}_1(u_y) \\ \tilde{L}_1(u_z) \end{pmatrix} = -\frac{\langle \cos^2\theta \rangle_{\mathrm{avg}}}{\sigma_t - g\sigma_s} \begin{pmatrix} i\,u_x \tilde{L}_0(u_x) \\ i\,u_y \tilde{L}_0(u_y) \\ i\,u_z \tilde{L}_0(u_z) \end{pmatrix} \ . \tag{7.15}$$

If the proportionality coefficient in this equation is independent of the position in $\boldsymbol{u}$-space, we can insert expressions for the Legendre polynomials expansion coefficients and inverse Fourier transform the equation, and the Fourier integrals and exponentials will cancel out. However, the mean square cosine of the

Fourier transformed radiance distribution $\langle\cos^2\theta\rangle_{\text{avg}}$, which is a part of the proportionality coefficient, may be a function of the position in $\boldsymbol{u}$-space. If this is the case, integrals and exponentials do not cancel out in the inverse Fourier transform. For the moment we will assume that $\langle\cos^2\theta\rangle_{\text{avg}}$ is independent of the position in $\boldsymbol{u}$-space. Under this assumption, we obtain

$$\begin{pmatrix} \int_{4\pi} \cos\alpha\, L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \\ \int_{4\pi} \cos\beta\, L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \\ \int_{4\pi} \cos\gamma\, L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \end{pmatrix} = -\frac{\langle\cos^2\theta\rangle_{\text{avg}}}{\sigma_t - g\sigma_s} \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x}\int_{4\pi} L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \\ \frac{\mathrm{d}}{\mathrm{d}y}\int_{4\pi} L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \\ \frac{\mathrm{d}}{\mathrm{d}z}\int_{4\pi} L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \end{pmatrix} .$$

A more elegant way to write this vector equation is

$$\int_{4\pi} \vec{\omega} L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega = -\frac{\langle\cos^2\theta\rangle_{\text{avg}}}{\sigma_t - g\sigma_s}\,\nabla\int_{4\pi} L(\boldsymbol{x},\vec{\omega})\,\mathrm{d}\omega \ ,$$

and this is exactly Fick's law of diffusion. The integrals are usually renamed as in Section 6.3 such that we have

$$\boldsymbol{F}(\boldsymbol{x}) = -D(\boldsymbol{x})\nabla\phi(\boldsymbol{x}) = -\frac{\langle\cos^2\theta\rangle_{\text{avg}}}{\sigma_t'(\boldsymbol{x})}\nabla\phi(\boldsymbol{x}) \ , \tag{7.16}$$

where $\boldsymbol{F}$ is the net flux (6.19), ϕ is the fluence (6.18), D is called the *diffusion coefficient*, and σ_t' is the *reduced extinction coefficient* defined by

$$\sigma_t' = (1-g)\sigma_s + \sigma_a = \sigma_t - g\sigma_s \ .$$

This derivation of Fick's law is more general than the derivations I have been able to find by previous authors. Fick [1855] originally derived this proportionality between net flux and the gradient of the fluence as a law for liquid diffusion. A derivation from the one-dimensional radiative transfer equation, which results in conclusions similar to ours (but for one-dimensional transport), has been presented by Glasstone and Edlund [1952].

The problem in generalisation of results derived for one-dimensional transport is that all equations are based on the assumption that radiance is rotationally invariant (axially symmetric) with respect to some given direction. The Fourier transform used in this section enables us to work around this problem. The idea of employing a Fourier transform for a generalised derivation of Fick's law was inspired by Waller's [1946] use of similar Fourier transforms (for a different purpose than deriving Fick's law). Fick's law has, of course, been derived for three-dimensional transport before [Case and Zweifel 1967; Ishimaru 1978], but these derivations are based on the P_1 approximation whereas the derivation given here is rigorous. In the P_1 approximation radiance $L(\boldsymbol{x},\vec{\omega})$ is approximated by the first two terms of a spherical harmonics expansion. This is a very restrictive assumption. There is substantial theoretical and experimental

evidence [Bothe 1941; Chandrasekhar 1950; Glasstone and Edlund 1952; Weinberg and Wigner 1958; Case and Zweifel 1967; Ishimaru 1989; Aronson and Corngold 1999; Elaloufi et al. 2003; Ripoll et al. 2005; Pierrat et al. 2006] that this assumption is only valid for nearly isotropic, almost non-absorbing media ($\sigma_s' >> \sigma_a$).

The *diffusion equation* follows quickly when Fick's law is available. Insertion of Fick's law (7.16) in Equation 6.21 gives the diffusion equation for a medium which is not an emitter:

$$\nabla \cdot (D(\boldsymbol{x}) \nabla \phi(\boldsymbol{x})) = \sigma_a \phi(\boldsymbol{x}) \ . \tag{7.17}$$

If sources were present in the medium, a few source terms would be present in this equation.

Before employing the diffusion equation, I would like to emphasise that it is only valid if $\langle \cos^2\theta \rangle_{\text{avg}}$ is independent of the position in $\boldsymbol{u}$-space. To determine when this is the case, we analyze the mean square cosine of the Fourier transformed radiance distribution using its definition (7.12). Since the third Legendre polynomial is $P_2(\cos\theta) = \frac{1}{2}(3\cos^2\theta - 1)$, we express cosine squared in terms of Legendre polynomials as follows:

$$\cos^2\theta = \frac{1}{3} P_0(\cos\theta) + \frac{2}{3} P_2(\cos\theta) \ . \tag{7.18}$$

After expansion of $\tilde{L}$ as a sum of Legendre polynomials, this result can be inserted in the definition of $\langle \cos^2\theta \rangle_{\text{avg}}$ and using the orthogonality of the Legendre polynomials, the result is

$$\langle \cos^2\theta \rangle_{\text{avg}} = \frac{1}{3} + \frac{2}{3} \frac{\tilde{L}_2(u)}{\tilde{L}_0(u)} \ . \tag{7.19}$$

In the P_1-approximation, $\tilde{L}_2$ vanishes and the diffusion equation (7.17) reduces to the traditional diffusion equation in which $D = 1/(3\sigma_t')$. However, as mentioned previously, the P_1-approximation is only valid for a very limited range of media. An important observation by Glasstone and Edlund [1952] is the following. If $\tilde{L}(\boldsymbol{u}, \vec{\omega})$ can be treated as a product of two functions, one of which depends on $\boldsymbol{u}$ and the other on $\vec{\omega}$, then all the expansion coefficients $\tilde{L}_n$ bear a constant ratio to $\tilde{L}_0$, and then Fick's law is valid. This condition is usually fulfilled in the asymptotic regions of a medium, that is, at considerable distances from sources and boundaries. It remains, then, to find an expression for the diffusion coefficient D which is valid for a broader range of media than those for which $\sigma_t' >> \sigma_a$.

From this point on, we can proceed in two different directions. One is due to Bothe [1941; 1942], the other to Waller [1946]. Bothe's approach is to insert

a trial solution in the radiative transfer equation (6.1), and derive an equation from which D can be determined. Waller's approach is to derive a continued fraction expression for $\tilde{L}_0$. With a considerable amount of algebraic manipulation, it is possible to transform this continued fraction into an expression for $L_0 = \phi$ [Holte 1948]. An expression for D can then be extracted from the expression for L_0 [Kuščer and McCormick 1991]. In both approaches, practically useful expressions are obtained by considering only a finite number of terms in a Legendre polynomials expansion of the phase function p.

A phase function which is very commonly used in graphics is the Henyey-Greenstein phase function [Henyey and Greenstein 1940]. It is a very convenient phase function to expand in Legendre polynomials because the coefficients are [Aronson and Corngold 1999]

$$p_n = g^n \ ,$$

where g is the asymmetry parameter (7.5). The full expansions by Bothe [1941; 1942] and Holte [1948] are reproduced and their range of validity is discussed by Aronson and Corngold [1999]. A practical expression for the diffusion coefficient D obtains if we truncate the original expression of Holte to first order:

$$D = \frac{1}{3\sigma_t'} \left(1 - \frac{4}{5} \frac{\sigma_a}{\sigma_s(1-g^2) + \sigma_a} \right)^{-1} \ . \tag{7.20}$$

This expression for the diffusion coefficient gives a much better result for absorbing media and media that exhibit anisotropic scattering [Ripoll et al. 2005]. This is important if we want to use diffusion rather than Monte Carlo path tracing because the larger the particles in a medium are, the more forward peaked (anisotropic) is the scattering of the medium [van de Hulst 1957]. For many natural materials the particles are big enough to exhibit highly anisotropic scattering. Most of the particles considered in Part IV are highly forward scattering.

7.2 Subsurface Scattering

Subsurface scattering is a name given to rendering algorithms that only consider the radiance emergent on the surface of a scattering medium. The first subsurface scattering model was presented by Blinn [1982]. The more general concept of subsurface scattering was formulated by Hanrahan and Krueger [1993]. The global formulation of subsurface scattering, that is, the formulation of subsurface scattering in terms of a BSSRDF was first considered by Pharr and Hanrahan [2000]. Except for Blinn's [1982] single scattering approximation, these models for subsurface scattering were based on Monte Carlo simulation and very expensive to evaluate.

To make volume rendering more practical, diffusion theory was introduced to graphics by Stam [1995]. He also suggests that diffusion theory might be used for subsurface scattering, but he notes the problem that diffusion theory is only valid in the asymptotic regions of a medium. The asymptotic region of a medium is some *mean free paths* from sources and boundaries of the medium. The mean free path is defined by $1/\sigma_t$ and it is the average distance that a ray of light penetrates into the medium before the first scattering event takes place. Disregarding the problem that diffusion theory is only strictly valid in the asymptotic regions, the diffusion theory was used by Jensen et al. [2001] to find a practical BSSRDF model for subsurface scattering. This is the model that people most often are referring to when they talk about subsurface scattering.

Subsurface scattering is currently one of the most popular rendering techniques. It is popular because it is a fast way of approximating the scattering that goes on beneath the surface of a medium. The more correct result is obtained by evaluation of the radiative transfer equation. To connect the theory of the previous chapters all the way to a practical rendering technique, I will give a quick overview of the subsurface scattering model introduced by Jensen et al. [2001].

To find a practical rendering technique, we would like to derive a BSSRDF for the rendering equation using diffusion theory. It is common to split the BSSRDF in two terms. One for single scattering $S^{(1)}$ and one for diffuse scattering S_d, such that

$$S = T_{12}(S_d + S^{(1)})T_{21} \ ,$$

where T_{12} and T_{21} are the Fresnel transmittance terms where the radiance enters and exits the medium, respectively. The advantage is that methods exist for evaluation of the single scattering part [Blinn 1982; Hanrahan and Krueger 1993]. Then we only need to worry about the diffuse part of the BSSRDF $S_d = S_d(|\boldsymbol{x}_i - \boldsymbol{x}_o|)$, where we assume that the diffuse part only depends on the distance between the point of incidence $\boldsymbol{x}_i$ and the point of emergence $\boldsymbol{x}_o$. This means that we can evaluate S_d by integrating some of the directional dependencies out of Equation 7.2.

Considering only the diffuse term and integrating over outgoing directions in Equation 7.2, the results is

$$\frac{\mathrm{d}M_d(\boldsymbol{x}_o)}{\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)} = \pi S_d(|\boldsymbol{x}_i - \boldsymbol{x}_o|) \ , \tag{7.21}$$

where $M_d = \mathrm{d}\Phi_o/\mathrm{d}A_o$ is called the *diffuse radiant exitance*. If we make the somewhat enervating assumption that no diffuse radiance is scattered at the surface in the inward direction, then the inward part of the integral defining the

diffuse net flux $\boldsymbol{F}_d$ will be zero at the surface, and we have

$$\vec{n} \cdot \boldsymbol{F}_d = \vec{n} \cdot \int_{2\pi} \vec{\omega} L(\boldsymbol{x}, \vec{\omega}) \, \mathrm{d}\omega = \int_{2\pi} L(\boldsymbol{x}, \vec{\omega}) \cos\theta \, \mathrm{d}\omega = M_d \ . \tag{7.22}$$

What is interesting here is that we now have a connection between the rendering equation and the diffusion theory.

The diffusion equation (7.17) is an expression which concerns the fluence ϕ. Many approximate analytical solutions exist in the literature for the diffusion equation. One such analytic solution is the *dipole approximation* [Farrell et al. 1992]:

$$\phi(\boldsymbol{x}) = \frac{\Phi}{4\pi D} \left(\frac{e^{-\sigma_{\mathrm{tr}} d_r}}{d_r} - \frac{e^{-\sigma_{\mathrm{tr}} d_v}}{d_v} \right) \ , \tag{7.23}$$

where Φ is the power of the dipole, $\sigma_{\mathrm{tr}} = \sqrt{3\sigma_a \sigma_t'}$ is the *effective transport coefficient*, $\sigma_t' = \sigma_a + (1-g)\sigma_s$ is the reduced extinction coefficient, g is the asymmetry parameter (7.5), and it is assumed that the medium is homogeneous such that the diffusion coefficient D is a constant. The dipole consists of two sources: a positive *r*eal source at the distance $d_r = |\boldsymbol{x}_r - \boldsymbol{x}_o|$ from the point where light exits the medium and a *v*irtual mirror source at the distance $d_v = |\boldsymbol{x}_v - \boldsymbol{x}_o|$ from the same point.

It is now only a matter of putting the math together. Inserting Fick's law of diffusion (7.16) in Equation 7.22, we have the following expression for the diffuse radiant exitance:

$$M_d = -D \ \vec{n} \cdot \nabla \phi(\boldsymbol{x}) \ .$$

Inserted in Equation 7.21, this gives an expression for the BSSRDF

$$\pi S_d(|\boldsymbol{x}_i - \boldsymbol{x}_o|) = -D \frac{\mathrm{d}(\vec{n} \cdot \nabla \phi(\boldsymbol{x}_o))}{\mathrm{d}\Phi_i(\boldsymbol{x}_i, \vec{\omega}_i)} \ .$$

To use this bridge between the rendering equation and diffusion theory, we insert the dipole approximation (7.23). If we orient the coordinate system such that the z-axis points in the direction of the inward normal at the point of incidence $\boldsymbol{x}_i$, the dipole sources lie on the z-axis and d_r and d_v are distance functions of z and $r = |\boldsymbol{x}_i - \boldsymbol{x}_o|$. See Figure 7.1. We get

$$\pi S_d(r) = -D \frac{\mathrm{d}\Phi}{\mathrm{d}\Phi_i} \frac{\partial}{\partial z} \left[\frac{1}{4\pi D} \left(\frac{e^{-\sigma_{\mathrm{tr}} d_r(z,r)}}{d_r(z,r)} - \frac{e^{-\sigma_{\mathrm{tr}} d_v(z,r)}}{d_v(z,r)} \right) \right] \ . \tag{7.24}$$

A few geometrical considerations lead to the following expressions for the distance functions:

$$d_r(z,r) = \sqrt{r^2 + (z + z_r)^2} \quad \text{and} \quad d_v(z,r) = \sqrt{r^2 + (z - z_v)^2} \ , \tag{7.25}$$

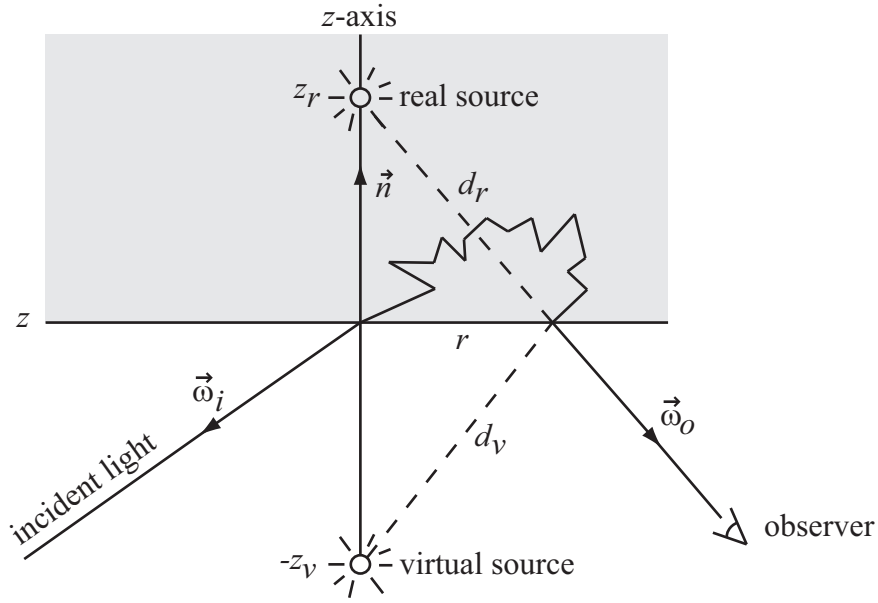


Figure 7.1: *The dipole approximation.*

where the signs have been chosen such that both z_r and z_v are positive. Inserting and differentiating in Equation 7.24 gives the following, when we subsequently set $z = 0$ to find S_d at the surface:

$$\pi S_d(r) = \frac{1}{4\pi}\frac{\mathrm{d}\Phi}{\mathrm{d}\Phi_i}\left(\frac{z_r(1+\sigma_{\mathrm{tr}}d_r)e^{-\sigma_{\mathrm{tr}}d_r}}{d_r^3} + \frac{z_v(1+\sigma_{\mathrm{tr}}d_v)e^{-\sigma_{\mathrm{tr}}d_v}}{d_v^3}\right) . \quad (7.26)$$

It remains to choose a power for the dipole and a position for each pole. Since the dipole represents light scattered diffusely within the medium, it makes sense to choose a power Φ which is proportional to the incident flux Φ_i. The diffuse color of the medium is approximately given by the scattering albedo $\alpha = \sigma_s/\sigma_t$ of the medium. Therefore we can use the reduced scattering albedo $\alpha' = \sigma_s(1-g)/\sigma_t'$, to denote the color of the dipole. The power of the dipole is then $\Phi = \alpha'\Phi_i$. An appropriate position for the real source is $z_r = 1/\sigma_t'$ which is one (reduced) mean free path below the point of incidence $\boldsymbol{x}_i$. The most obvious choice would then be the same z_v, but a trick is used here to correct for the error introduced by the assumption that no diffuse radiance is scattered at the surface in the inward direction. This assumption is not true if the medium has an index of refraction different from that of air (because then some diffuse radiance will be reflected back into the medium at the surface). A position for the virtual source taking

this internal reflection into account is $z_v = z_r + 2AD$ [Moulton 1990; Farrell et al. 1992], where the parameter A is related to the diffuse Fresnel reflectance R_{dr} as follows [Groenhuis et al. 1983]:

$$A = \frac{1 + R_{\mathrm{dr}}}{1 - R_{\mathrm{dr}}} \ . \tag{7.27}$$

The diffuse Fresnel reflectance is defined by the integral over the hemisphere of the cosine weighted Fresnel reflectance. An approximate empirical formula is [Egan and Hilgeman 1979]:

$$R_{\mathrm{dr}} = -1.440n'^{-2} + 0.710n'^{-1} + 0.668 + 0.0636n' \ , \tag{7.28}$$

where $n' = n'_{\mathrm{from}}/n'_{\mathrm{to}}$ is the relative index of refraction, that is, the ratio of the refractive index for the medium containing the incident light ray to that of the medium which the light is incident upon. In our case, since we consider internal reflection, $n' = n'_{\mathrm{inside}}/n'_{\mathrm{outside}}$.

If we insert these expressions for the power of the dipole Φ and the distances to the poles z_r and z_d in Equation 7.26, we get the following practical BSSRDF for the diffuse scattering:

$$S_d(r) = \frac{\alpha'}{4\pi^2}\left(\frac{(1+\sigma_{\mathrm{tr}}d_r)e^{-\sigma_{\mathrm{tr}}d_r}}{\sigma'_t d_r^3} + \left(\frac{1}{\sigma'_t} + 2AD\right)\frac{(1+\sigma_{\mathrm{tr}}d_v)e^{-\sigma_{\mathrm{tr}}d_v}}{d_v^3}\right) \ ,$$

where

$$\begin{aligned} d_r &= \sqrt{r^2 + (1/\sigma'_t)^2} \\ d_v &= \sqrt{r^2 + (1/\sigma'_t + 2AD)^2} \ . \end{aligned}$$

The parameter A is given by Equation 7.27 and D is the diffusion coefficient which we gave a practical approximation for in Equation 7.20. Furthermore, we recall that $\sigma'_t = \sigma_a + (1-g)\sigma_s$ is the reduced extinction coefficient and $\sigma_{\mathrm{tr}} = \sqrt{3\sigma_a\sigma'_t}$ is the effective transport coefficient. Thus we see that the practical BSSRDF of Jensen et al. [2001] is completely described by the index of refraction of the surrounding medium, which is usually air ($n'_{\mathrm{outside}} = 1$), and the macroscopic optical properties of the material. These are, in summary,

- n' - the real part of the index of refraction
- σ_a - the absorption coefficient
- σ_s - the scattering coefficient
- g - the asymmetry parameter.

Part II of this thesis explores how it is possible to compute the macroscopic optical properties from a description of the particle composition of a material.

7.3 Conclusions

This part of the thesis has covered a broad range of theories concerning the propagation of light. To get started, a historical study of theories of light was presented (Chapter 2). This work is, to my knowledge, the first in-depth, source-based historical study that pinpoints events with relevance for graphics in the development of theories of light. Besides demonstrating that we can use the old theories of light to construct fast rendering techniques (Section 2.4), one of the key observations was that the development of realistic image synthesis, to a certain extent, has followed the development of theories of light. To investigate how we might continue this development, the chapters which followed the historical perspective strived to connect the theories of light currently used in graphics all the way up to quantum electrodynamics.

The connection between the theories of light used in graphics and the theory of quantum electrodynamics is important because it reveals in what aspects the behaviour of light has been simplified in graphics. Even more importantly, it reveals how the material properties used in graphics connect to the more microscopic properties of materials measured by physicists. The relationship between microscopic and macroscopic material properties is further investigated in Part II. The theory of quantum electrodynamics was only described introductorily (Chapter 3). In my opinion, it is important to have an introductory understanding of the true nature of light. The first purpose of the introductory description was to formulate a renderer which captures all known details in the behaviour of light (Section 3.4). While such a renderer, in general, is too computationally expensive, it provides the information we need to formulate more macroscopic rendering algorithms which capture phenomena such as diffraction. The introductory description of quantum electrodynamics (Chapter 3) serves other purposes as well: It provides the means to understand the absorption and emission properties of materials (this subject is investigated introductorily in Chapter 8); it introduces the concept of operators to handle problems of high dimensionality (we will return to such operators in Part III, where we will look at multidimensional appearance models from a geometrical point of view); finally, it demonstrates how the quantum theory of light simplifies to Maxwell's electromagnetic field theory.

The electromagnetic field theory is essential in many aspects of this thesis. The description of the theory (Chapter 4) was a review with emphasis on the parts of the theory that are important for the thesis. In particular, the origin of the complex index of refraction was discussed. This is important because the imaginary part of the index of refraction denotes absorption, and the correct handling of absorbing media is a general theme throughout the thesis. Since the complex index of refraction is not considered so often in the literature, some effort was

spent to show that the formulae from the electromagnetic theory which we use most often in graphics (the law of reflection, the law of refraction, the Fresnel equations) are also valid for media with a complex index of refraction. Furthermore, the flow of energy in an electromagnetic field was considered because the flow of energy determines the intensity of the colours that we see, and, hence, it is what we would like to trace in order to render realistic images.

To trace the flow of energy in an electromagnetic field, geometrical optics is used in graphics. Geometrical optics is a simplification of the wave theory of light to a ray theory of light. It has been described many times in the literature, but usually it is described for homogeneous waves in a heterogeneous material, or for inhomogeneous waves in a homogeneous material. The description given in this thesis (Chapter 5) is also valid for inhomogeneous waves in a heterogeneous medium. The reason for this general treatment of geometrical optics is that waves are almost always inhomogeneous when they propagate through an absorbing medium. Absorption is important in graphics as it tells us a lot about the appearance of a material. The presented treatment of geometrical optics involved a new idea for real-valued ray tracing of heterogeneous, absorbing media. Theoretically, the new idea more faithfully follows the true flow of energy through an absorbing medium. If the absorption is strong, one might object that we do not need to trace rays through the medium as the energy will quickly be absorbed. Being able to trace rays through an absorbing particle is on the other hand very valuable. It enables us to determine the scattering of an absorbing particle of arbitrary shape using simple ray tracing. In Chapter 9 we will see that it is not so easy to compute the scattering of a particle in an electromagnetic field using an analytical approach. It remains to be investigated how accurately the new ray tracing scheme models the true electromagnetic scattering by particles.

In graphics we model a material composed of many particles using macroscopic scattering properties and radiative transfer theory. After the description of the path that rays of light follow through a medium, the connection between scattering of electromagnetic waves by particles and the scattering properties used in radiative transfer theory was given ample attention (Chapter 6). This is a difficult subject because radiative transfer theory was not originally based on the physical theories of light. The presented treatment of the subject emphasised the limiting assumptions that we make when we go to a macroscopic description of scattering and use it with radiative transfer theory. Finally, we connected the presented theory to rendering. This was done by a short description of Monte Carlo path tracing for rendering volumes (Section 6.4).

Since volume rendering is an expensive computation, we often only consider light emergent on surfaces. To complete the connection between physical theories of light and the rendering algorithms used in practice, the connection to

surface-based rendering techniques was described in this chapter. To describe how radiative transfer connects to diffusion-based rendering techniques (and to find out how general the connection is), a new and more general derivation of Fick's law of diffusion was also presented (Section 7.1). Finally, it was demonstrated how the very popular rendering technique called subsurface scattering (the dipole approximation) is derived from diffusion theory.

In summary, this part covered the historical development leading to quantum electrodynamics and the simpler theory of light currently employed in graphics; the simplification of quantum electrodynamics to the theory of light that we use; and various considerations over the role of macroscopic optical material properties in the theories of light and in different rendering techniques. Throughout this part, I have strived to make the theory general such that it covers light propagation in a large group of the materials that we find in the real world. This is important in a graphics context, where we would like to capture the appearance of nature rather than the behaviour of light in an artificial setup such as a system of lenses.

Part II

MATTER

Chapter 8

Electron Theory

"It was most suggestive," said Holmes. "It has long been an axiom of mine that the little things are infinitely the most important."

Sir Arthur Connan Doyle, from *A Case of Identity*

Although the first part of this thesis is mainly about theories of light, it also contains much information about matter. Indeed we have introduced several parameters to describe the properties of materials at a macroscopic level. What we have not discussed is how to obtain these optical properties. The easy way out is to say that we can simply measure them, but that would leave us with an incredibly large amount of work to do. We cannot measure the properties of every material that we want to use. Even if we make a huge database of material properties, we would often need a slightly different version of a material or the properties of a material under different conditions. Then we could adjust the measured properties manually, but unfortunately some of the properties are not very intuitive. It is, for example, difficult to guess how the scattering properties of an iceberg changes when its temperature changes, or how milk scatters light differently when its fat content decreases. We could also measure these dependencies, but it makes the measurements even more difficult and our database would increase with an incredible amount of tabulated data. In short, it is not practical to measure optical properties at too macroscopic a level.

Even if it is too complicated to derive optical properties of materials from the quantum theories, it is very valuable to know how they connect to the more microscopic properties of matter. The smaller the level at which we model the

properties, the fewer "primitives" do we have. At a more microscopic level the database is smaller, and it becomes easier to model the relationship between the properties and the physical conditions of the material. At an atomic level the motion of the atoms tells us how the material relates to its temperature. If we look at the particle composition of a material (here referring to light scattering particles, not quantum particles), we will be able to model how the optical properties change when we change the contents of the material relative to each other. In the following we want to find out from which point of departure it makes sense to calculate the optical properties. And at what level should we use measurements? The chapters of this part explore the relations between electrons and the index of refraction (Chapter 8); between the index of refraction and the scattering of a particle (Chapter 9); between the properties of a single particle and the properties of a cloud of particles (Chapter 10); and between the spectral optical properties and the trichromatic colour-values used in a rendering (Chapter 11).

The first material properties we discussed in Part I were the charge and current densities in the microscopic Maxwell equations (4.1–4.4). As described in Chapter 3, they model the mean effect of the electrons in a charge field. All the electrons are however not alike. They have different definite and angular momenta. The possible states of these momenta depend on the nature of the material. The polarisation and magnetisation vectors described in Section 4.2 are ways of describing the properties of electrons which are bound to atoms. The polarisation vector $\boldsymbol{P}$ is the dipole moment per unit volume which approximates an assembly of atoms. The magnetisation vector $\boldsymbol{M}$ is the net magnetic moment per unit volume [Feynman et al. 1964, Sec. 35-4]. For N atoms with average magnetic moment $\boldsymbol{\mu}_{\text{avg}}$, we have

$$\boldsymbol{M} = N\boldsymbol{\mu}_{\text{avg}} \ .$$

To explain what magnetic moment is, we need to know what magnetism is. It is caused by orbit motion and spin of electrons in atoms. For quantum mechanical reasons the momentum of the orbit motion around the nucleus is half the angular momentum which describes the spin of the electron [Feynman et al. 1964, Sec. 34-2]. Thus only one momentum is needed to describe the orbit and spin. This momentum is called the *magnetic momentum* $\boldsymbol{\mu}$ and it is the charge $-q_e$ of the electron divided by its mass m times the angular momentum $\boldsymbol{J}$:

$$\boldsymbol{\mu} = -\frac{q_e}{m}\boldsymbol{J} \ .$$

As the next step we introduced the electric susceptibility (or polarisability) χ_e and the magnetic susceptibility χ_m. The former assumes a proportionality relation between the polarisation vector $\boldsymbol{P}$ and the electric field vector $\boldsymbol{E}$ such

that $\boldsymbol{P} = \varepsilon_0 \chi_e \boldsymbol{E}$. The latter assumes proportionality between the magnetisation vector $\boldsymbol{M}$ and the magnetic vector $\boldsymbol{H}$ such that $\boldsymbol{M} = \chi_m \boldsymbol{H}$. Based on these assumptions, as well as the assumption that the density of free currents is proportional to the electric field vector, we arrived at the isotropic material equations (4.27–4.29). The proportionality factors appearing in these equations were used to define the index of refraction n (cf. Equation 4.41). The index of refraction is the first of the optical properties which we use extensively in graphics. The relation of the index of refraction to the more microscopic descriptions of matter is completely obscured by the intermediate material properties. In the following section we will see if we can get a better understanding of the refractive index.

8.1 The Index of Refraction

The relationship between electrons and the index of refraction is not intuitively clear when we look at the description which is based on the speed of light in vacuum c, the permittivity ε, the permeability μ, and the conductivity σ of the material:

$$n = c\sqrt{\mu(\varepsilon + i\sigma/\omega)} \ .$$

It is, however, not so difficult to explain what the index of refraction means (approximately) when we consider a plane wave as the solution of Maxwell's equations. The real part describes the ratio of the speed of light to the phase velocity of the wave $n' \approx c/v$, while the imaginary part describes the absorption of light by the material $n'' \approx \sigma_a \lambda / 4\pi$ (cf. Section 4.3).

Since n' is greater than one for almost all materials (under rare circumstances it may also be less than one), it looks as if light slows down when entering a material, but we know that photons *always* move at the speed of light. The reason for this peculiarity is that it is not the same photons emerging from the material. The emerging photons have been emitted anew by the electrons in the material. This is also the case for the reflected photons. Howcome they emerge in the reflected and refracted directions? Because, when we add up the probability amplitudes for the photons to go in all the different possible directions through space, it so happens that the amplitude is significantly larger for the photons to move in the reflected and refracted directions. The phase velocity (and therefore n') must then be determined by the ability of the material to absorp and reemit photons.

Let us limit this discussion of the index of refraction to non-magnetic materials ($\mu = 1$). The motion of the charges in a material defines its ability to absorp and emit photons. The degree to which an incident light field sets the charges

in motion is described by the electric susceptibility χ_e of the material. If we let $\alpha(\omega)$ denote the (frequency dependent) degree to which an atom is electrically susceptible, the *atomic polarisability*, we have [Feynman et al. 1964, Sec. 32-3]

$$\boldsymbol{P} = \varepsilon_0 N \alpha(\omega) \boldsymbol{E}_{\text{local}} \ , \tag{8.1}$$

where N is the number of atoms per unit volume and $\boldsymbol{E}_{\text{local}}$ is the electric field at a single atom.

If we take the macroscopic Maxell equations (4.20–4.23), and assume that there are no free currents or charges (and still that the material is non-magnetic such that $\boldsymbol{M} = \boldsymbol{0}$), some algebraic manipulation gives

$$\nabla^2 \boldsymbol{E} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = -\frac{1}{\varepsilon_0}\nabla(\nabla \cdot \boldsymbol{P}) + \frac{1}{\varepsilon_0 c^2}\frac{\partial^2 \boldsymbol{P}}{\partial t^2} \ .$$

Assuming that the materials are isotropic, the divergence of the polarisation vector is zero. If we also assume that the waves are plane, the polarisation vector will also behave as a plane wave as it is proportional to $\boldsymbol{E}$. Then the equation reduces to

$$-k^2 \boldsymbol{E} + \frac{\omega^2}{c^2}\boldsymbol{E} = -\frac{\omega^2}{\varepsilon_0 c^2}\boldsymbol{P} \ . \tag{8.2}$$

Thinking of the atom as being spherical (which is alright for most liquids and for atoms in a cubic crystal lattice), the local electric field is [Feynman et al. 1964, Sec. 11-4]

$$\boldsymbol{E}_{\text{local}} = \boldsymbol{E} + \frac{1}{3\varepsilon_0}\boldsymbol{P} \ .$$

With the expression for the polarisation vector (8.1), this expression for the local field gives

$$\boldsymbol{P} = \frac{3\varepsilon_0 N\alpha}{3 - N\alpha}\boldsymbol{E} \ ,$$

which we insert in Equation 8.2 to get the following expression for the index of refraction:

$$n^2 = 1 + \frac{3N\alpha}{3 - N\alpha} \ .$$

Another way to write it is

$$3\frac{n^2 - 1}{n^2 + 2} = N\alpha \ . \tag{8.3}$$

This equation is known sometimes as the Clausius-Mossotti equation, sometimes as the Lorentz-Lorenz formula. It was derived first in a slightly different form by Mossotti [1850] and Clausius [1879] who assumed that atoms are small conducting spheres. Lorentz [1880] and Lorenz [1880] derived this relation between the index of refraction and the atomic polarisability as we see it here.

There could be several different atoms in a material. To capture this, we rewrite the left-hand side of Equation 8.3 such that it is a sum over the number densities for the different atoms:

$$3\frac{n^2-1}{n^2+2} = \sum_j N_j \alpha_j(\omega) \ . \tag{8.4}$$

Under all these simplifying assumptions (non-magnetic, isotropic material of simple structure) we are able to approximate the index of refraction by calculating the number density for the different atoms in a material. These number densities are found in a simple way using the chemical formula for the material and Avogadro's number.

As Newton found out in his experiments with prisms, the index of refraction is different for different wavelengths. This is because the atomic polarisability α depends on frequency. Since atoms work as damped oscillators with several resonant frequencies, they follow the equation [Feynman et al. 1964, Sec. 32-1]

$$\alpha(\omega) = \frac{q_e^2}{\varepsilon_0 m} \sum_k \frac{f_k}{\omega_{0k}^2 - \omega^2 + i\gamma_k \omega} \ , \tag{8.5}$$

where the atom acts as if it has resonators of charge $f_k q_e$, mass $f_k m$, and damping coefficient (or dissipation) γ_k at the natural frequency ω_{0k} of oscillation mode k. To find these coefficients that determine the wavelength dependency ($\omega = 2\pi c/\lambda$) of the index of refraction, we have to employ quantum mechanical concepts. For simple atoms it is possible to derive values for the coefficients using the Hamiltonian operator described in Chapter 3, but in many cases we will still have to rely on a number of measurements. Sometimes only a few oscillation modes are necessary to describe the index of refraction in the visible part of the spectrum (and note that that includes the absorption spectrum). This means that we have a theory from which we are able to approximate the index of refraction using only a limited number of constants for each atom. This makes the size of the database we need manageable.

The formula for the atomic polarisability (8.5) reveals that the real and imaginary parts of the index of refraction depend on the same set of constants. Thus there is also an internal relation between them. There are some limits to the combinations of n' and n'' which occur in nature. Absorption affects the phase velocity of light in a medium and vice versa. If we insert the expression for the atomic polarisability in the Clausius-Mossotti equation (8.4), and isolate the real and imaginary parts of the index of refraction, the result is not pretty. All the coefficients influence both the real and imaginary parts. The damping coefficient is, however, far more significant in absorption (the imaginary part) than in the real part of the refractive index.

Absorption occurs when a photon is annihilated in the process of lifting an electron to a higher energy state. Conversely, an emission occurs when a photon is created (or liberated) as an electron drops to a lower energy state. There is consequently a relationship between absorption, emission, and the coefficients needed to compute the index of refraction. In the following section we will briefly explore these relations.

8.2 Absorption and Emission

Suppose we have a material with a number density N_1 of atoms in energy state E_1 and a number density N_2 of atoms in energy state E_2. We let the difference between the energy states be such that the absorption of one photon will lift an atom from E_1 to E_2 and emission of a photon will do the opposite. This means that $E_2 - E_1 = \hbar\omega$. The Boltzmann law then says that in thermal equilibrium:

$$N_2 = N_1 e^{-\frac{E_2 - E_1}{kT}} = N_1 e^{-\frac{\hbar\omega}{kT}} \ , \tag{8.6}$$

where $k = 1.38 \cdot 10^{-23}\,\mathrm{J/K}$ is the boltzmann constant. Letting $|N_p\rangle$ denote a state with N_p photons of frequency ω and using the photon creation and annihilation operators, $\hat{a}^\dagger$ and $\hat{a}$, we recall from Section 3.1 that the probability amplitude of a photon being created (i.e. emitted) is given by

$$\hat{a}^\dagger |N_p\rangle = \sqrt{N_p + 1}\, |N_p + 1\rangle \ ,$$

while the probability amplitude of a photon being annihilated (i.e. absorbed) is

$$\hat{a} |N_p\rangle = \sqrt{N_p}\, |N_p - 1\rangle \ .$$

The rate at which photons are created and annihilated is the sum of the probabilties for all the photons present. At thermal equilibrium the rate of emission must equal the rate of absorption. Therefore, if we recall that the probabilities are the square of the probability amplitudes, we get

$$N_1 N_p = N_2 (N_p + 1) \ .$$

Using the ratio N_2/N_1 as given by Boltzmann's law (8.6), we get the number density of photons in energy state $\hbar\omega$:

$$N_p = \frac{1}{e^{\frac{\hbar\omega}{kT}} - 1} \ . \tag{8.7}$$

If we find the different oscillation modes of the atoms in a material, we will be able to say what the number density of photons is at thermal equilibrium for

each mode. In other words, we can use this result to derive the emission and absorption spectra of materials.

We could now proceed to find the probabilities of electrons to shift from one energy state to another using the Hamiltonian operator described in Chapter 3, but it would take us too far afield. Instead, we simply note a few results from classical physics. If $\rho(\omega)\,\mathrm{d}\omega$ denotes the energy per unit volume of radiation in the angular frequency interval $[\omega, \omega + \mathrm{d}\omega]$, then at thermal equilibrium we have [Milonni 1994]

$$\rho(\omega_0)\,\mathrm{d}\omega = \frac{\omega_0^2}{\pi^2 c^3} U\,\mathrm{d}\omega \ ,$$

where U is the radiation energy per unit volume. If we consider a material with atoms that are able to radiate in all possible modes, we can replace ω_0 by ω in this formula. Such a material is called a blackbody, and if we use our result from before to find the radiation energy $U = \hbar\omega N_p$, we get the blackbody emission spectrum using Equation 8.7:

$$\rho(\omega)\,\mathrm{d}\omega = \frac{\omega^2}{\pi^2 c^3} \frac{\hbar\omega}{e^{\frac{\hbar\omega}{kT}} - 1}\,\mathrm{d}\omega \ .$$

Using different relations between the quantities ($\omega = 2\pi c/\lambda$, $\hbar = 2\pi h$), we can also write it in terms of wavelengths:

$$\rho(\lambda)\,\mathrm{d}\lambda = \rho(\omega)\left|\frac{\mathrm{d}\omega}{\mathrm{d}\lambda}\right|\mathrm{d}\lambda = \frac{8\pi c h \lambda^{-5}}{e^{\frac{hc}{\lambda kT}} - 1}\,\mathrm{d}\lambda \ .$$

The blackbody emission spectrum has been used frequently in graphics to model light sources. For sources that do not fit the blackbody emission spectrum, we use tabulated emission properties. In this chapter I have tried to show that there are other options. If the material is not approximately a blackbody, we have the option to look at the resonant frequencies (ω_{0k}) of the atoms in the material. It is important to note that these resonant frequencies (or natural modes of oscillation) in atoms control all the macroscopic material properties. We can use them to compute both emission spectra and indices of refraction (which includes absorption spectra). Ultimately the connection between graphics and these physical theories can make us able to approximate the appearance of a material by looking at their chemical formulae. This opens up for many new applications of graphics. By visualizing the different components in a material each on their own and in different concentrations, we will be able learn how the appearance of materials relate to their chemical composition. This is useful, for example, if we want to predict or design the appearance of a material.

To continue this construction of a bridge between microscopic and macroscopic material properties, we will relate the indices of refraction to the scattering properties of materials in the following chapters.

Chapter 9

Particles as Spheres

MOYERS: *Perfection* would *be a bore, wouldn't it?*

CAMPBELL: *It would have to be. It would be inhumane. The umbilical point, the humanity, the thing that makes you humane and not supernatural and immortal — that's what's lovable.*

Joseph Campbell and Bill Moyers, from *The Power of Myth*

In Chapter 6 we introduced macroscopic scattering properties by considering the scattering of a particle embedded in a host medium. The scattering was described without considering the actual geometry of the particles. Instead, we let an undetermined scattering matrix $\mathbf{S}(\theta, \phi)$ describe the scattering of an arbitrary particle. Using this matrix, we derived a scattering cross section and a phase function for a particle. Finally, we combined the scattering cross sections with the number density of particles at different sizes to get the scattering coefficient and phase function which we use in realistic rendering. In this chapter we will show that it is possible to compute the scattering matrix if we assume that the particles are perfect spheres (Sec. 9.1). The formulae that we need to evaluate the scattering matrix are surprisingly complicated. So we will also discuss how one should evaluate the formal solution in practice (Sec. 9.2). Finally, we will briefly consider how to handle non-spherical particles (Sec. 9.3).

The calculation of scattering by a perfect sphere has been considered for more than a century. The formulae were originally derived by Lorenz [1890], and formally they have not changed much since. In the original theory, Lorenz

considered light from a transparent medium scattered by a transparent sphere. He used the ratio $N = n'_{\text{med}}/n'_p$ of the real refractive index of the host medium to that of the spherical particle. Mie [1908] derived the same formulae over again using Maxwell's equations. Mie was considering colloidal suspensions of metallic particles. Since metals have a very significant imaginary part in their index of refraction, he generalised the original formulae using a complex index of refraction n_p for the spheres. If you think about what types of matter that you are able to model with the original formulae, they are extremely limited. Water drops in air is an example which makes us able to model atmospheric phenomena such as clouds, mist, haze, and rainbows. It is difficult to think of other types of matter consisting of transparent particles in a transparent host medium. So, while the theoretical contribution of Mie was small, his extension of the theory was a considerable improvement in terms of materials that we are able to model. The theory of scattering by spherical particles is today called the Lorenz-Mie theory. Kerker [1969, Sec. 3.4] gives an excellent historical review of the different contributions to the theory in its initial development (and of how little the different contributors knew of each other's work).

Even with the extension of Mie, the class of materials that we are able to model as possibly absorbing spheres in a transparent host is very limited. There are many particles which are approximately spheres, but it is rarely the case that the particles are embedded in a transparent host. The blue colour of the seas is caused by the weak absorption of water. Hence, even water is not transparent if the volume we are considering is large enough. Examples of absorbing spheres in a transparent host are then materials such as paints and plastics, suspensions of metallic particles in a transparent solvent, coloured glass, and the like. It would be much preferable if we were able to model the general case where the refractive index of the host medium is also allowed to be complex. This extension of the Lorenz-Mie theory was first considered by Mundy et al. [1974]. They used formulae formally equivalent to those of Lorenz and Mie, but they swapped the real-valued n'_{med} in the final formulae for the complex version n_{med} and this is how an absorbing host medium has subsequently been modelled.

After the paper by Mundy et al. [1974] there has been much discussion on scattering by particles in an absorbing host. The discussion is concerned with the problem explained in Section 6.1: we cannot really consider the scattering cross section to be an independent property of the material because there is an exponential attenuation term in the direction towards the observer left over in the formula (6.10). It means that we cannot really tell how much light was scattered by the particle if we look at the light that reaches a far away observer. This problem has not yet been solved. Another problem which people have neglected is that the Lorenz-Mie formulae were derived for homogeneous waves of light. In Chapters 4 and 5 we saw that waves are only very rarely homogeneous when propagating in an absorbing medium. Belokopytov and Vasil'ev [2006] have de-

rived the scattering of plane inhomogeneous waves by an absorbing particle in a non-absorbing host. This is a very special case because waves are as rarely inhomogeneous in a transparent medium as they are rarely homogeneous in an absorbing medium. Our goal in this chapter is to find the Lorenz-Mie formulae for the scattering of an absorbing sphere in an absorbing host. To do it properly, we need to take inhomogeneous waves in an absorbing medium into account. Even if we choose to limit ourselves to the more traditional case of homogeneous waves in our implementation, we should, at least theoretically, find out what difference it makes when the waves are inhomogeneous.

9.1 Scattering by a Sphere

Consider a plane wave scattered by a spherical particle embedded in a host medium. We take it that both the host medium and the particle are isotropic, homogeneous substances. Otherwise plane waves would not be a good approximation (cf. Section 4.3). Note that we make no assumptions about the absorption properties of the host medium and the particle, and we do not require the plane wave to be homogeneous or inhomogeneous. This treatment is therefore more general than previous work on the subject. However, we do this only as a theoretical exercise. I have not yet tried out the general theory in practice.

In the following we work with all the field vectors in their time-free form (this is acceptable as the exponential term involving time cancels out in the time-harmonic Maxwell equations). Furthermore we denote the complex refractive indices of the host medium and the particle n_{med} and n_p respectively. We are aiming at an expression for the scattered field. All we know at the moment (cf. Section 6.1) is that in the far field we can think of the scattered field as a spherical wave. To describe the relation between this spherical wave and the incident wave, we used a scattering matrix $\mathbf{S}(\theta, \phi)$. The scattering matrix relates the $\parallel$-polarised and the $\perp$-polarised components of the incident electric field to the same components of the Poynting vector of the scattered light (cf. Equations 6.8 and 6.9). Thus if we find the components of the scattering matrix, we also have an expression for the scattered field.

In Chapter 6 (Figure 6.1) we chose our coordinate system such that light is incident along the z-axis. This means that the z-axis points along the direction of the time-averaged Poynting vector for the incident field. For homogeneous plane waves the direction of the Poynting vector for the incident field is parallel to both the real and imaginary parts of the wave vector, but in general that is not necessarily the case. This means that we ought to think of the wave vector $\boldsymbol{k}$ as having two different directions: one for the real part $\boldsymbol{k}'$ and one for the

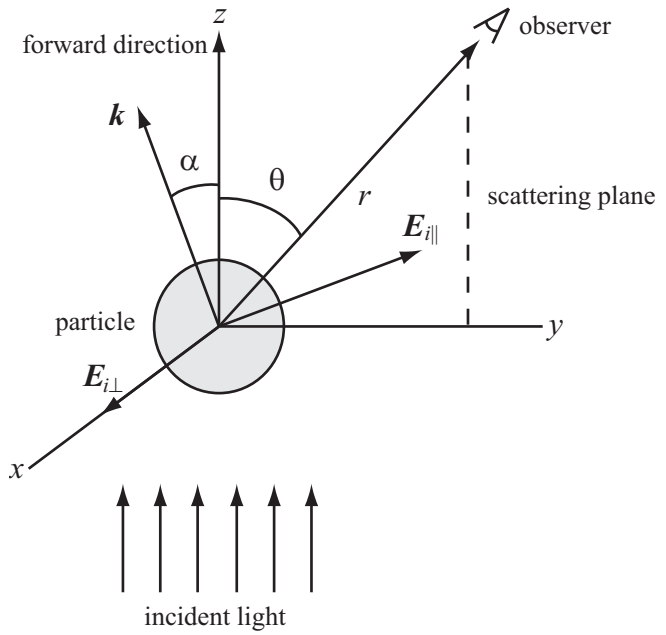


Figure 9.1: *Scattering of light by a spherical particle in an absorbing host.*

imaginary part $\boldsymbol{k}''$. However, let us still think of it as if it were real, and only worry about the imaginary part when necessary.

The scattering plane is spanned by the z-axis and the direction of the scattered light. As we did in Chapter 5, we may identify TE components with $\perp$-polarised components and TM components with $\|$-polarised components. The $\perp$-polarised component of the electric field vector is by definition perpendicular to the scattering plane. The $\|$-polarised component lies in the scattering plane. For a homogeneous wave, the direction of $\boldsymbol{E}_{i\|}$ would also be perpendicular to the z-axis, but that is not true in general.

Since we are working with a sphere, it is of no consequence with respect to particle geometry which way we choose to orient the x and y axes of the coordinate system. So let us choose that the x-axis is in the direction of the $\perp$-polarised component of the electric vector. By definition the $\|$-polarised component lies in the yz-plane, and we have

$$\boldsymbol{E}_{i0} = E_{i\perp}\vec{e}_x + E_{i\|}(a\vec{e}_y + b\vec{e}_z) \ , \tag{9.1}$$

where $\vec{e}_x$, $\vec{e}_y$, and $\vec{e}_z$ are the unit vectors in the direction of the axes of the coordinate system, $E_{i\perp}$ and $E_{i\|}$ are the complex amplitudes of the two polarisation

components of $\boldsymbol{E}_{i0}$, and a and b are the direction cosines of $\boldsymbol{E}_{i0}$ in the y and z directions respectively. See Figure 9.1.

Using that the $\|$-polarised component is also transverse magnetic (TM), we have the following general relation between the (complex) magnetic vector $\boldsymbol{H}$ and the $\|$-polarised component of the electric vector $\boldsymbol{E}_{\|}$ (it was obtained by modifying a case described by Born and Wolf [1999, §11.4.1]):

$$E_y^{\|} = -\frac{c\mu}{ik_0 n^2}\frac{\partial H_x}{\partial z} \quad , \quad E_z^{\|} = \frac{c\mu}{ik_0 n^2}\frac{\partial H_x}{\partial y} \ ,$$

where c is the speed of light in vacuum, μ is the permeability, and $k_0 = \omega/c$ is the wave number in vacuum. Dividing by $E = \sqrt{\boldsymbol{E}\cdot\boldsymbol{E}}$, gives the direction cosines. From the plane wave Maxwell equations (4.35–4.38) and the definition of the index of refraction (4.41), we have

$$H_0 = E_0 \frac{n}{c\mu} \ ,$$

and $H_x = H$ because $H_y = H_z = 0$ for the TM component.

Using these general relations with our incident light, we get

$$a = -\frac{k_z}{k_0 n_{\text{med}}} = -\cos\alpha \quad , \quad b = \frac{k_y}{k_0 n_{\text{med}}} = \sin\alpha \ , \tag{9.2}$$

where α is a complex angle denoting the direction of the wave vector in the scattering plane (see again Figure 9.1).

From the plane wave Maxwell equations (in particular Equation 4.37), we know that $\boldsymbol{k}\cdot\boldsymbol{E}_0 = 0$. Calculating this dot product using Equations 9.1 and 9.2, we get

$$\boldsymbol{k}\cdot\boldsymbol{E}_0 = k_x E_{i\perp} - k_y E_{i\|}\frac{k_z}{k_0 n_{\text{med}}} + k_z E_{i\|}\frac{k_y}{k_0 n_{\text{med}}} = k_x E_{i\perp} \ .$$

Therefore we have $k_x = 0$ as long as there is a TE component. This means that both the real and imaginary part of the wave vector $\boldsymbol{k}$ lie in the scattering plane. Consequently, the electric vector of the incident wave has the following structure:

$$\boldsymbol{E}_i = \boldsymbol{E}_0 e^{i(k_y y + k_z z)} \ .$$

This is quite different from the case of a homogeneous wave, where only one coordinate is left in the exponential. For the inhomogeneous wave *the scattering is not symmetric around the forward direction.* A result which is very counter-intuitive since we are working with a sphere. It means that we can use only part of the results of Lorenz and Mie (and many others) for inhomogeneous waves. To find a solution, we look to the work of Belokopytov and Vasil'ev [2006].

Suppose we introduce the following slightly unusual set of spherical coordinates:

$$\begin{aligned} x &= \cos\beta \\ y &= \sin\beta\sin\theta \\ z &= \sin\beta\cos\theta \ , \end{aligned}$$

where the x-axis is the polar axis such that the scattering plane is given by $\sin\beta = 1$ and the forward direction when also $\theta = 0$. The angle θ is shown in Figure 9.1. With these spherical coordinates, the expression for our incident field is

$$\boldsymbol{E}_i = \boldsymbol{E}_0 e^{ir(k_y \sin\beta\sin\theta + k_z \sin\beta\cos\theta)} \ .$$

Using the complex angle α, we can also write it as

$$\boldsymbol{E}_i = \boldsymbol{E}_0 e^{ikr\sin\beta(\sin\alpha\sin\theta+\cos\alpha\cos\theta)} = \boldsymbol{E}_0 e^{ikr\sin\beta\cos(\theta-\alpha)} \ , \qquad (9.3)$$

where $k = k_0 n_{\text{med}}$. In fact we are only interested in the scattering plane so we could omit the $\sin\beta$ factor in the exponential. The version including $\sin\beta$ is very close to the form of incident wave for which Belokopytov and Vasil'ev found a solution. The generalisation we have made is in the definition of α (which is purely imaginary in their treatment). If you have an inhomogeneous wave in a non-absorbing medium, you have $\boldsymbol{k}\cdot\boldsymbol{k} = k_0^2 n_{\text{med}}^2$ to be a real number. Assuming k_z real and k_y purely imaginary in our Equations, we would have an inhomogeneous wave in a non-absorbing medium, and in that case the structure of our incident wave would be equivalent to the one addressed by Belokopytov and Vasil'ev [2006]. Thus the generalisation is a bit like Mie's generalisation of Lorenz' results: it is perhaps a small theoretical contribution, but it significantly extents the range of naturally occurring phenomena which the theory is able to model.

The spherical harmonics expansion based on the associated Legendre polynomials P_n^m, which was found by Belokopytov and Vasil'ev [2006] for this type of exponential function (9.3), is

$$e^{ikr\sin\beta\cos(\theta-\alpha)} = \sum_{n=0}^{\infty} i^n (2n+1)\frac{\psi_n(kr)}{kr} \sum_{m=0}^{n} R_n^m P_n^m(\cos\beta)\cos(m(\theta-\alpha))$$

with

$$R_n^m = \begin{cases} (-1)^{(n-m)/2}\, 2^{1-n-\delta_{mn}} \frac{(n+m)!}{((n-m)/2)!((n+m)/2)!} & \text{for } n+m = 2l \\ 0 & \text{for } n+m = 2l+1 \end{cases} \ ,$$

where $n, m \in \{0, 1, 2, \dots\}$ and $l \in \{1, 2, \dots\}$, and δ_{nm} is the Kronecker delta. The function ψ is a Riccati-Bessel function which we will return to later. The wave number $k = k_0 n_{\text{med}}$ is complex in our case, but that does not invalidate the result.

What we are really interested in is the scattered wave. In particular we would like to determine the four components of the scattering matrix $\mathbf{S}(\theta, \phi)$. It has been shown many times in the literature [Debye 1909; van de Hulst 1957; Kerker 1969] that we do not need the expression for all the radial components of the scattered field to find the scattering matrix. We only need the Debye potentials (which are related to the scalar potentials discussed in Section 4.1). In the expansion of Belokopytov and Vasil'ev [2006] the Debye potentials for the scattered wave outside the sphere are:

$$u = \frac{1}{k^2}\sum_{n=1}^{\infty} i^{n-1}\frac{2n+1}{n(n+1)}a_n\zeta_n(kr)U_n(\theta,\beta) \tag{9.4}$$

$$v = \frac{1}{k^2}\sum_{n=1}^{\infty} i^{n-1}\frac{2n+1}{n(n+1)}b_n\zeta_n(kr)V_n(\theta,\beta) \ , \tag{9.5}$$

where

$$\begin{aligned}
V_n(\theta,\beta) &= E_{i\parallel}V_{1n}(\theta,\beta) + E_{i\perp}V_{2n}(\theta,\beta)\\
U_n(\theta,\beta) &= -E_{i\perp}V_{1n}(\theta,\beta) + E_{i\parallel}V_{2n}(\theta,\beta)\\
V_{1n}(\theta,\beta) &= \sum_{m=1}^{n} R_n^m m P_n^m(\cos\beta)\sin(m(\theta-\alpha))\\
V_{2n}(\theta,\beta) &= \sum_{m=1}^{n} R_n^m (n-m) P_n^m(\cos\beta)\cos(m(\theta-\alpha)) \ .
\end{aligned}$$

The spherical function $\zeta_n(z)$, where z is a complex number, is composed of the spherical Bessel functions j_n and y_n such that

$$\zeta_n(z) = z(j_n(z) + iy_n(z)) \ .$$

The coefficients a_n and b_n fortunately turn out to be the traditional Lorenz-Mie coefficients when we go to the far field. We will return to those shortly. In the far field, we get expressions for the components in the scattering matrix (cf. Equation 6.8). As we already know, $S_3 = S_4 = 0$ for perfect spheres [van de Hulst 1957], the remaining two components of the scattering matrix are [Belokopytov and Vasil'ev 2006]

$$S_1 = \sum_{n=1}^{\infty}\frac{2n+1}{n(n+1)}\left(a_n\frac{\partial U_n(\theta,\beta)}{\partial\beta} + b_n\frac{1}{\sin\beta}\frac{\partial V_n(\theta,\beta)}{\partial\theta}\right) \tag{9.6}$$

$$S_2 = \sum_{n=1}^{\infty}\frac{2n+1}{n(n+1)}\left(a_n\frac{1}{\sin\beta}\frac{\partial U_n(\theta,\beta)}{\partial\theta} - b_n\frac{\partial V_n(\theta,\beta)}{\partial\beta}\right) \ . \tag{9.7}$$

This is the solution for the scattering of an inhomogeneous plane wave by an absorbing sphere in an absorbing host. It clearly demonstrates how unwillingly

the mathematics we know describe the scattering of electromagnetic waves by particles.

Let us try to interpret the angles appearing in this generalised formulation of the Lorenz-Mie theory. We chose that the x-axis is along the direction of the $\perp$-polarised component of the electric vector. The angle β is the polar angle between this component and the scattering plane. This is *always* a right angle. Therefore we have $\beta = 90°$. The azimuth angle θ is the angle in the scattering plane between the forward direction and the direction of the scattered light. This means that θ is the standard scattering angle. Another angle in the formulae is the complex angle α. It is a measure of the inhomogeneity of the incident wave, and it is given by the direction cosines of the wave vector $\boldsymbol{k}$ of the incident light (cf. Equation 9.2).

If we were to render a medium with a phase function derived from the scattering functions presented here, it is new that we would have to keep track of the wave vector. It is, however, easily done for homogeneous media (i.e. media with the same macroscopic optical properties throughout). It is easy because in homogeneous media we have $\nabla\mathscr{S} = \boldsymbol{k}$, and we found a formula for computing the gradient of the optical path $\nabla\mathscr{S}$ in Chapter 5 (cf. Equation 5.34).

It should be mentioned that for a homogeneous wave, we get $k_z = k$ and $k_y = 0$, and then $\alpha = 0$. This simplifies the structure of the exponential term in the expression for the incident wave, and we get the components of the scattering matrix in the commonly available form [Lorenz 1890; Mie 1908; van de Hulst 1957; Kerker 1969]:

$$S_1(\theta) = \sum_{n=1}^{\infty} \frac{2n+1}{n(n+1)} \left(a_n \pi_n(\cos\theta) + b_n \tau_n(\cos\theta)\right) \tag{9.8}$$

$$S_2(\theta) = \sum_{n=1}^{\infty} \frac{2n+1}{n(n+1)} \left(a_n \tau_n(\cos\theta) + b_n \pi_n(\cos\theta)\right) \quad , \tag{9.9}$$

where the functions π_n and τ_n are related to the Legendre polynomials P_n as follows:

$$\pi_n(\cos\theta) = \frac{P_n^1(\cos\theta)}{\sin\theta} = \frac{\mathrm{d}P_n(\cos\theta)}{\mathrm{d}(\cos\theta)}$$

$$\tau_n(\cos\theta) = \frac{\mathrm{d}P_n^1(\cos\theta)}{\mathrm{d}\theta} = \cos\theta\pi_n(\cos\theta) - \sin^2\theta \frac{\mathrm{d}\pi_n(\cos\theta)}{\mathrm{d}(\cos\theta)} \quad .$$

Their numeric evaluation can be found in standard references on Lorenz-Mie theory [Dave 1969; Bohren and Huffman 1983].

It remains in this discussion of scattering by a spherical particle to give expressions for the Lorenz-Mie coefficients a_n and b_n. In the far field they are given

by [Lorenz 1890]

$$a_n = \frac{n_{\text{med}}\psi_n'(y)\psi_n(x) - n_p\psi_n(y)\psi_n'(x)}{n_{\text{med}}\psi_n'(y)\zeta_n(x) - n_p\psi_n(y)\zeta_n'(x)} \tag{9.10}$$

$$b_n = \frac{n_p\psi_n'(y)\psi_n(x) - n_{\text{med}}\psi_n(y)\psi_n'(x)}{n_p\psi_n'(y)\zeta_n(x) - n_{\text{med}}\psi_n(y)\zeta_n'(x)} \ , \tag{9.11}$$

where the primes $'$ denote derivative. The spherical functions $\psi_n(z)$ and $\zeta_n(z)$ are known as Riccati-Bessel functions. They are related to the spherical Bessel functions $j_n(z)$ and $y_n(z)$ as follows:

$$\begin{aligned} \psi_n(z) &= z j_n(z) \\ \zeta_n(z) &= z(j_n(z) - i y_n(z)) \ . \end{aligned}$$

The argument z is an arbitrary complex number, the arguments x and y used for the Lorenz-Mie coefficients are related to particle and host medium as follows:

$$x = \frac{2\pi r n_{\text{med}}}{\lambda} \quad \text{and} \quad y = \frac{2\pi r n_p}{\lambda} \ ,$$

where λ is the wavelength in vacuum and r is the radius of the spherical particle.

When computers came around, it turned out to be quite difficult to find a numerically stable way of evaluating the spherical functions ψ_n and ζ_n for complex arguments. Eventually, the numerical difficulties were solved for complex y [Kattawar and Plass 1967; Dave 1969; Wiscombe 1980]. This is sufficient for the traditional Lorenz-Mie theory with a non-absorbing host medium. When people started considering spheres in an absorbing host, starting with Mundy et al. [1974], it became necessary to find a robust way of evaluating the Lorenz-Mie coefficients for complex x as well. This is considerably more difficult. The following section describes a robust evaluation scheme proposed by Frisvad et al. [2007].

9.2 Evaluating Lorenz-Mie Coefficients

In the case of an absorbing host medium n_{med} has an imaginary part and then the parameter x, used for evaluation of the Lorenz-Mie coefficients (9.10–9.11), is complex. The consequence is that most numerical evaluation schemes become unstable because the Riccati-Bessel functions enter the exponential domain and run out of bounds. Previously this instability has been solved for complex y parameters (absorbing particles), in the following we will describe a scheme for robust evaluation when x is also complex.

To avoid the ill-conditioning of the Riccati-Bessel functions ψ_n and ζ_n, the Lorenz-Mie coefficients are rewritten in a form involving only ratios between them [Kattawar and Plass 1967]

$$a_n = \frac{\psi_n(x)}{\zeta_n(x)} \frac{n_{\text{med}} A_n(y) - n_p A_n(x)}{n_{\text{med}} A_n(y) - n_p B_n(x)} \tag{9.12}$$

$$b_n = \frac{\psi_n(x)}{\zeta_n(x)} \frac{n_p A_n(y) - n_{\text{med}} A_n(x)}{n_p A_n(y) - n_{\text{med}} B_n(x)} \quad . \tag{9.13}$$

Here $A_n(z)$ and $B_n(z)$ denote the logarithmic derivatives of $\psi_n(z)$ and $\zeta_n(z)$ respectively:

$$A_n(z) = \frac{\psi_n'(z)}{\psi_n(z)} \quad \text{and} \quad B_n(z) = \frac{\zeta_n'(z)}{\zeta_n(z)} \quad .$$

The ratio A_n is only numerically stable with downward recurrence. Therefore the following formula is employed for its evaluation [Kattawar and Plass 1967]

$$A_n(z) = \frac{n+1}{z} - \left(\frac{n+1}{z} + A_{n+1}(z) \right)^{-1} \quad . \tag{9.14}$$

This formula is also valid for the ratio B_n, but then it is unfortunately unstable for both upward and downward recurrence [Cachorro and Salcedo 1991]. Instead, we use a different formula for B_n which has been developed by Mackowski et al. [1990] in the field of multilayered particles embedded in a non-absorbing medium. It is numerically stable with upward recurrence for any complex argument [Mackowski et al. 1990]:

$$B_n(z) = A_n(z) + \frac{i}{\psi_n(z)\zeta_n(z)} \tag{9.15}$$

$$\psi_n(z)\zeta_n(z) = \psi_{n-1}(z)\zeta_{n-1}(z) \left(\frac{n}{z} - A_{n-1}(z) \right) \left(\frac{n}{z} - B_{n-1}(z) \right) \quad . \tag{9.16}$$

It remains to give a recurrence relation for the ratio $\psi_n(z)/\zeta_n(z)$ in Equations 9.12 and 9.13. Recent developments in the context of multilayered particles, provide a recurrence relation that works well for small $\text{Im}(z)$ [Wu and Wang 1991; Yang 2003]:

$$\frac{\psi_n(z)}{\zeta_n(z)} = \frac{\psi_{n-1}(z)}{\zeta_{n-1}(z)} \frac{B_n(z) + n/z}{A_n(z) + n/z} \quad . \tag{9.17}$$

The restriction to small $\text{Im}(z)$ is not a problem in graphics applications, as a larger $\text{Im}(z)$ means that the host medium is highly absorbing, and then we would not be able to make out the effect of particle scattering anyway.

The amplitude functions (9.8–9.9) are defined by an infinite sum, and in order to get a decent approximation, we must find an appropriate number of terms M to

sum. This is also necessary for initialisation of the downward recurrence (9.14) which computes $A_n(x)$ and $A_n(y)$. A formula determining M, which has both an empirical [Wiscombe 1980; Mackowski et al. 1990] and a theoretical [Cachorro and Salcedo 1991] justification, is

$$M = \left\lceil |x| + p|x|^{1/3} + 1 \right\rceil \quad , \tag{9.18}$$

where $p = 4.3$ gives a maximum error of 10^{-8}. It is possible to calculate an approximate initial value for the downward recurrence (9.14), but, as explained by Dave [1969], the recurrence is not sensitive to the initial value, and therefore we can arbitrarily choose $A_M(z) = 0$.

Once $A_0(z), \ldots, A_M(z)$ have been computed for both $z = x$ and $z = y$, we are able to find the ratios $B_n(x)$ and $\psi_n(x)/\zeta_n(x)$ as well as the Lorenz-Mie coefficients, a_n and b_n, step by step. Note that there is no need to store $B_n(x)$ and $\psi_n(x)/\zeta_n(x)$ since they are computed using upward recurrences. These recurrences should be initialised by

$$\begin{aligned} B_0(z) &= i \\ \psi_0(z)\zeta_0(z) &= \tfrac{1}{2}\left(1 - e^{i2z}\right) \\ \psi_0(z)/\zeta_0(z) &= \tfrac{1}{2}\left(1 - e^{-i2z}\right) \quad . \end{aligned} \tag{9.19}$$

Recall that there is a direct relationship between wavelength λ and the size parameters x and y. This tells us that the Lorenz-Mie coefficients are spectrally dependent and should preferably be sampled at different wavelengths. They are also dependent on the particle radius r and are valid for spherical particles of arbitrary size as long as they do not exhibit diffuse reflection (which is only possible if the particle size greatly exceeds the wavelength and even so, the surface of the particle might still be smooth) [van de Hulst 1957]. Furthermore the equations provided in this section reveal that the complex refractive index of each particle inclusion, as well as that of the host medium, are needed as input parameters for computing the optical properties of a scattering material.

Having a robust way to compute the Lorenz-Mie coefficients, makes us able to evaluate the components of the scattering matrix. Either we can use the traditional approach, Equations 9.8 and 9.9, which is valid for homogeneous waves, or we can use the more general approach, Equations 9.6 and 9.7, which is also able to handle inhomogeneous waves. When we have obtained the scattering amplitudes, we are able to find the extinction and scattering cross sections as well as the phase function of the particle. These are all well defined quantities for particles in a non-absorbing medium. For a particle in an absorbing medium, we saw in Chapter 6 that the scattering cross section is a problematic quantity because the resulting formula depends on the distance to the observer.

When particles are embedded in an absorbing host, the extinction cross section C_t is the only well defined observable quantity [Bohren and Gilra 1979]. It is computed using an optical theorem first presented by van de Hulst [1949; 1957]. The original theorem by van de Hulst is valid for particles of arbitrary shape and size, but it only applies to a non-absorbing host medium. To account for an absorbing host, we use a slightly modified equation presented by Bohren and Gilra [1979]:

$$C_t = 4\pi \mathrm{Re}\left(\frac{S(0)}{k^2}\right) \quad , \tag{9.20}$$

where $S(0) = S_1(0) = S_2(0)$ is the amplitude in the forward direction of the scattered wave and $k = 2\pi n_{\mathrm{med}}/\lambda$ is the wave number. Since the host medium was assumed by van de Hulst to be non-absorbing, n_{med} and therefore also k were assumed real and moved outside the Re operator (which takes the real part of a complex number). This is not allowed if the host medium is absorbing as the result would be a meaningless complex extinction coefficient. Correction by discarding the imaginary part of the result would not be a good approximation (except when particle absorption is considerably stronger than that of the host medium [Bohren and Gilra 1979]). Inserting the expression for $S(0)$ in this optical theorem (9.20) we get

$$C_t = \frac{\lambda^2}{2\pi} \sum_{n=1}^{\infty} (2n+1) \mathrm{Re}\left(\frac{a_n + b_n}{n_{\mathrm{med}}^2}\right) \quad . \tag{9.21}$$

A form has not been found for the scattering cross section C_s which is independent of the distance to the observer, but we still have to approximate C_s to evaluate the radiative transfer equation (6.1). We use a far-field approximation which has been reported to be consistent with measured data [Randrianalisoa et al. 2006; Yin and Pilon 2006]. The chosen formula is identical to the scattering cross section for transparent media except for two correction terms: an exponential term and a geometrical term γ. The formula is

$$C_s = \frac{\lambda^2 e^{-4\pi r \,\mathrm{Im}(n_{\mathrm{med}})/\lambda}}{2\pi\gamma |n_{\mathrm{med}}|^2} \sum_{n=1}^{\infty} (2n+1) \left(|a_n|^2 + |b_n|^2\right) \, , \tag{9.22}$$

where r in the exponential term is the uncertain part of the equation because it ought to be the distance to where the scattered wave is observed. This distance is unknown, and consequently it has been projected to the particle surface, such that r denotes the particle radius.

The geometrical term γ accounts for the fact that the incident wave changes over the surface of the particle as a consequence of the absorbing host medium. It is defined by [Mundy et al. 1974]

$$\gamma = \frac{2(1 + (\alpha - 1)e^{\alpha})}{\alpha^2} \quad , \tag{9.23}$$

where $\alpha = 4\pi r \,\mathrm{Im}(n_{\mathrm{med}})/\lambda$ and $\gamma \to 1$ for $\alpha \to 0$. Note that α is 0 when the medium is transparent and close to 0 for small particles in a weakly absorbing medium. To avoid numerical errors, one should use $\gamma = 1$ for $\alpha < 10^{-6}$.

The precision of the far field approximation (9.22,9.23) has recently been reviewed [Fu and Sun 2006] and compared to experimental data [Randrianalisoa et al. 2006; Yin and Pilon 2006]. The conclusion is that it (as expected) does not give entirely accurate results, but it does give physically plausible results. It is also concluded that significant errors can result if the absorption of the host medium is ignored (this is especially true when the size parameter x is large).

The expression for the phase function and the asymmetry parameter are the same in transparent and absorbing media [Yang et al. 2002] (except for the fact that waves are inhomogeneous in absorbing media). The phase function for unpolarised light is [van de Hulst 1957]

$$p(\theta) = \frac{|S_1(\theta)|^2 + |S_2(\theta)|^2}{4\pi \sum_{n=1}^{\infty} (2n+1)\left(|a_n|^2 + |b_n|^2\right)} \quad .$$

Some authors fail to specify that this is the phase function for unpolarised light. For linearly polarised light it is:

$$p(\theta, \varphi) = \frac{|S_1(\cos\theta)|^2 \sin^2\varphi + |S_2(\cos\theta)|^2 \cos^2\varphi}{2\pi \sum_{n=1}^{\infty} (2n+1)\left(|a_n|^2 + |b_n|^2\right)} \quad .$$

For both unpolarised and linearly polarised light the asymmetry parameter (which is defined by the integral of the cosine weighted phase function over all solid angles) is [van de Hulst 1957]:

$$g = \frac{\sum_{n=1}^{\infty} \left\{ \frac{n(n+2)}{n+1} \mathrm{Re}(a_n a_{n+1}^* + b_n b_{n+1}^*) + \frac{2n+1}{n(n+1)} \mathrm{Re}(a_n b_n^*) \right\}}{\frac{1}{2} \sum_{n=1}^{\infty} (2n+1)\left(|a_n|^2 + |b_n|^2\right)}, \tag{9.24}$$

where the asterisks * denote the complex conjugate.

This concludes the robust scheme for computing the optical properties of a sphere in a host medium. In the next section we will take a brief look at scattering by a non-spherical particle.

9.3 Non-Spherical Particles

Lorenz-Mie theory is for spherical particles, but particles are not always spherical. Suppose we have a cylindrically shaped particle. Then we could go through

all the lengthy calculations of Section 9.1 using cylindrical coordinates instead of spherical coordinates and arrive at a different formal solution. With that in hand, we could try to make the calculations practical and general as we did for the sphere in Section 9.2. If the expressions for the sphere had been simple and beautiful, it might be tempting, but the solution for the sphere is not at all nice. There do exist a number of theories for scattering of other perfect mathematical shapes like cylinders, hexagonal columns and plates, etc. They are useful because some materials actually do consist of particles approximately of these shapes. Halos are, for example, the result of scattering by hexagonal ice crystals in the atmosphere. Instead of the mathematical approach, let us use a more practical approach to non-spherical particles.

With the theory we have already developed, we are able to handle non-spherical particles in two ways: one option is to approximate a non-spherical particle by an appropriate collection of spherical particles, another option is to use the ray tracing approach described in Chapter 5. Let us look at the former option first.

It is not obvious what set of spheres we should choose to model a non-spherical particle in the best way. Many different concepts have been tried: Equal-volume spheres, equal-area spheres, etc. The best approach I am aware of is that of Grenfell and Warren [1999]. They use volume-to-area equivalent spheres. As opposed to equal-volume and equal-area spheres, the volume-to-area equivalent spheres have proven to be quite exact. They have been tested for cylinders [Grenfell and Warren 1999], hexagonal columns and plates [Neshyba et al. 2003], and hollow columns and plates [Grenfell et al. 2005]. In most cases the error is less than 5%. At least this is true for scattering and extinction coefficients. The approximation is, as could be expected, less accurate with respect to the phase function.

To represent a particle of volume V and surface area A by a collection of spheres, the radius of the equivalent spheres is found simply using the volume to surface area ratio of a sphere [Grenfell and Warren 1999]:

$$r_{\mathrm{eq}} = 3\frac{V}{A} \quad . \tag{9.25}$$

Since the number of equivalent spheres is not equal to the number of non-spherical particles, the number density must be adjusted accordingly [Grenfell and Warren 1999]:

$$\frac{N_{\mathrm{eq}}}{N} = \frac{3V}{4\pi r_{\mathrm{eq}}^3} \quad . \tag{9.26}$$

The equivalent radius r_{eq} and the equivalent number density N_{eq} are then used for computing the macroscopic optical properties (see Section 6.2 or the next chapter) with Lorenz-Mie theory for computing the cross sections of the equiv-

alent spheres. This is a simple and practical approach which gives rather nice results.

The alternative approach is to use the ray tracing described in Chapter 5. This approach is a particularly interesting from a graphics perspective. To make it work, we need a practical way of storing the result of the ray traced scattering. A cube map is a way of constructing a directional function using six rendered images. The cube map was proposed by Greene [1986]. Suppose we place the particle in an axis-aligned bounding box, and illuminate it using a directional light source in the direction of the z-axis. Then if we render the six faces of the bounding box (looking inward at the particle), and store them in a cube map, we have a numerical representation of the scattered light in all directions around the particle. This means that the cube map will contain the magnitude of the scattered Poynting vector $|\boldsymbol{S}_s|$ in all directions. If we integrate it over all directions on the unit sphere, we get the scattering cross section C_s. If we then divide the values in the cube map by the scattering cross section, we get the phase function of the particle.

This approach demonstrates the ambiguity of the particle being embedded in an absorbing host. How do we handle it? From where do we attenuate the incident light, from where do we render the images? The distance to the particle matters, and in principle the incident light is thought of as coming from infinitely far away. Likewise the scattered light is thought of as being observed from infinitely far away. The best way to solve this problem is probably to ray trace a spherical volume containing a collection particles embedded in the host medium. This corresponds to the shell functions proposed by Moon et al. [2007]. If the shell function is captured using the ray tracing approach described in Chapter 5, I believe it would be a very faithful way of capturing the scattering properties of a medium. It also solves the problem that non-spherical particles should be oriented in a random fashion and that many different particle sizes should be present. In a very direct manner the shell function will give us the macroscopic scattering properties of the material. It will, however, be more expensive to compute than the approach based on Lorenz-Mie theory (since fairly many particles might be needed to construct a spherical volume representative of the material).

With this suggestion of using shell functions to capture the combined or the *bulk* optical properties of a material, we have ventured into the subject of the next chapter. In the next chapter we will consider the quick approach to computing bulk optical properties. We will use the single particle properties, which we now know how to find using Lorenz-Mie theory, to find macroscopic scattering properties. It is common to use a simplified phase function because we have to evaluate it many times during a rendering. In the next chapter we will also look at macroscopic or simplified representations of the phase function.

Chapter 10

Bulk Optical Properties

For verily not by design did the first beginnings of things station themselves each in its right place guided by keen intelligence, nor did they bargain sooth to say what motions each should assume, but because many in number and shifting about in many ways throughout the universe they are driven and tormented by blows during infinite time past, after trying motions and unions of every kind at length they fall into arrangements such as those out of which this our sum of things has been formed

Lucretius (c. 99 – 55 B.C.), from *On the Nature of Things*

The subject of going from microscopic to macroscopic scattering properties has already been discussed in Section 6.2. Using the somewhat doubtful independent scattering approximation (6.15), we arrived at a simple way of converting the scattering cross sections of specific particles to the bulk scattering coefficient σ_s of a medium (6.16). The number densities of the different sizes and types of particles in the medium are key ingredients in this conversion. Therefore we look, in this chapter, at a few examples of number density distributions which are often found in natural materials (Sec. 10.1).

When we consider the particles embedded in a medium from a macroscopic point of view, we not only want bulk scattering properties, but also the bulk absorption and extinction of the medium. In the previous chapter we saw that the computations leading to the scattering cross section also lead to the extinction cross section of the particles. This means that number density distributions also give us a way of computing the bulk extinction coefficient σ_t. Knowing the

bulk scattering and extinction coefficients, we also know the bulk absorption coefficient ($\sigma_a = \sigma_t - \sigma_s$). The bulk absorption coefficient leads to the imaginary part of the bulk index of refraction (cf. Equation 4.43). This poses a challenge because, as mentioned in Section 8.1, there is a relation between the real and imaginary parts of the refractive index. To uphold a physically plausible set of material properties, the real part should be changed according to the new imaginary part of the bulk index of refraction. An approximate formula finding the real part of the bulk index of refraction for particles in a non-absorbing host has been derived by van de Hulst [1957]. We generalise this formula such that it also works for particles in an absorbing host (Sec. 10.1).

The phase function is used frequently in a rendering based on the radiative transfer equation. Therefore it should be either pre-computed and stored in a look-up table or simple to evaluate. The Lorenz-Mie formulae are not simple to evaluate, but many simplified expressions for the phase function are based on the asymmetry parameter g. As noted in Section 9.2, the Lorenz-Mie coefficients also gives a way of finding the asymmetry parameter for a particle. When phase functions for many different particles are combined to form one phase function describing the bulk directional scattering properties of a medium, it is referred to as the *ensemble* phase function. Likewise the combined asymmetry parameter is called the *ensemble* asymmetry parameter. Conclusively, in this chapter, we investigate how one should go about computing the ensemble phase function and the ensemble asymmetry parameter (Sec. 10.2).

10.1 Number Density Distributions

If we assume that extinction by one particles is independent of the extinction by another, in the same way as we assume that scattering is independent, the bulk extinction and scattering coefficients are given by

$$\sigma_j(\boldsymbol{x}) = \int_0^\infty C_j(\boldsymbol{x}, r) N(r) \, \mathrm{d}r \ , \tag{10.1}$$

where $\boldsymbol{x}$ is the position in the medium, r is the radius of a particle, $N(r)$ is the particle number density distribution, and j is either t referring to extinction or s referring to scattering. Of course, the integral disappears if the particles are all the same size. But in most natural materials a single particle radius cannot predict the optical properties correctly. The integral will, however, always be zero outside some limited interval $[r_{\min}, r_{\max}]$ of particle sizes.

Particle *size distribution* is the common term for distributions that we can use to find the number densities of particles of different sizes. One type of size dis-

tribution, which is often encountered in the literature, is the *volume frequency distribution* $r^3N(r)$. Such distributions typically follow a log-normal distribution. Log-normal distributions are often described by a mean particle size μ and a coefficient of variation $c_v = \sigma/\mu$, where σ is the standard deviation.

If we find that the volume frequency of some type of particle in a medium follows the log-normal distribution with mean value μ and standard deviation σ, the volume frequency distribution is given by

$$r^3N(r) = \frac{1}{r\beta\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln r-\alpha}{\beta}\right)^2} \quad , \tag{10.2}$$

where r is the particle radius and

$$\alpha = \ln\mu - \frac{1}{2}\ln\left(\frac{\sigma^2}{\mu^2}+1\right) \quad \text{and} \quad \beta = \sqrt{\ln\left(\frac{\sigma^2}{\mu^2}+1\right)} \quad .$$

This type of distribution is commonly observed for small particles. Distributions of larger particles tend to follow a *power law*. If the particle number density of some type of particle in a medium follows a power law, the distribution is given by

$$N(r) = N_* r^{-\alpha} \quad , \tag{10.3}$$

where α is usually determined empirically and N_* is a constant which is determined by the relationship between number density and the volume fraction of the medium occupied by the considered type of particle.

Now we have an idea about how to find the scattering and extinction due to one type of particle in a medium. Let us see how to deal with several different types of particles. In the following we let A denote the set of homogeneous substances appearing as particles in a host medium. Measured data are sometimes available which specify the volume fraction v_i, $i \in A$, of each particle inclusion that is present in the bulk medium. Otherwise this is a reasonable choice of input parameters. The number density distribution $N(r)$ specifies the number of particles per unit volume with radii in the interval $[r, r + \mathrm{d}r]$. This means that the volume fraction occupied by a particle inclusion, which consists of spherical particles, is

$$v = \frac{4\pi}{3}\int_{r_{\min}}^{r_{\max}} r^3N(r)\,\mathrm{d}r \quad . \tag{10.4}$$

Suppose we measure the particle size distributions for some sample of material. Then we would have empirical functions or tabulated data that fit the volume fractions of the particles in the original sample. Most probably the original volume fractions are not the volume fractions we desire in our medium. Equation 10.4 is important because it explains how we find the original volume fraction

$v_{\text{original},i}$ of particle type i. If the volume fraction v_i is desired rather than $v_{\text{original},i}$, the measured number density should be scaled by $v_i/v_{\text{original},i}$.

Because we assume that particles scatter light independently, not only scattering cross sections are additive, but also scattering coefficients (and extinction coefficients) are additive. We let $\sigma_{s,i}$ and $\sigma_{t,i}$ denote the scattering and extinction coefficients for every individual particle inclusion $i \in A$ in the considered medium. Finding the bulk scattering coefficient is straightforward:

$$\sigma_s = \sum_{i \in A} \sigma_{s,i} \quad . \tag{10.5}$$

Note that volume fractions are *not* included in this formula, because they are a part of the number density distributions.

In a transparent medium, the extinction coefficient is defined by an equivalent sum, but in an absorbing medium an important correction must be made. Since the host medium is a part of the extinction process, a non-absorbing particle will reduce the extinction of the bulk medium. This means that the extinction cross sections can be negative [Bohren and Gilra 1979]. The extinction cross section resulting from the Lorenz-Mie theory is, in other words, relative to the absorption of the host medium and the necessary correction is to include the host medium absorption in the sum. For this purpose, we compute the bulk extinction coefficient for particles in an absorbing medium by

$$\sigma_t = \sigma_{a,\text{med}} + \sum_{i \in A} \sigma_{t,i} \quad , \tag{10.6}$$

and the bulk absorption coefficient is given by the simple relation $\sigma_a = \sigma_t - \sigma_s$. These bulk coefficients are never negative.

To compute the refractive index of the bulk medium, we follow van de Hulst's [1957] derivation of a formula for the effective index of refraction, but we remove the assumptions of non-absorbing media and particles of only one radius. This gives the following approximate relation for the real part of the bulk refractive index:

$$\text{Re}(n_{\text{bulk}}(\lambda)) = \text{Re}(n_{\text{med}}(\lambda)) + \lambda \sum_{i \in A} \int_0^\infty \text{Im}\left(\frac{S_{i,r,\lambda}(0)}{k^2}\right) N_i(r)\, \mathrm{d}r \quad , \tag{10.7}$$

where $S_{i,r,\lambda}(0)$ is the amplitude in the forward direction of a wave of wavelength λ scattered by a particle of radius r and type $i \in A$, $N_i(r)$ is the number density distribution, and k is the wave number. The imaginary part of the effective index of refraction is not the correct imaginary part for the bulk medium, but rather a term related to the total extinction of the medium. The correct imaginary part is found by its relation to the bulk absorption coefficient (4.43 with $\cos\theta = 1$).

This concludes our discussion of all the bulk optical properties of a medium except the phase function which is the subject of the next section. One of the reasons why we would like to be able to compute optical properties is that it makes it easier to find the relation between the optical properties and the physical conditions of the medium. In this context, it should be mentioned that the number density distributions, or size distributions, as well as the radii of the particles may depend on the physical conditions of the medium. For example temperature T and pressure P. The scattering and extinction cross sections depend on the refractive indices of the particles n_i, $i \in A$, and that of the host medium n_{med}. These indices of refraction may again depend on T and P. Measurements that describe these dependencies are sometimes available in the literature. With such measurements, we can make models that compute the optical properties of a medium as a function of the physical conditions of the medium.

10.2 Macroscopic Phase Functions

The asymmetry parameter g and phase function p are normalised properties related to the amount of scattering by every particle. Say we denote the asymmetry parameter of a single particle $g_p(r)$ and the corresponding phase function $p_p(r)$, where r is the particle radius. The ensemble asymmetry parameter g_i and phase function p_i that combine all the different sizes of particle type i, are then found by a weighted average, where the weights are the associated scattering cross sections. For particle inclusion i we have

$$p_i(\theta) = \frac{1}{\sigma_{s,i}} \int_{r_{\min}}^{r_{\max}} C_{s,i}(r) p_{p,i}(\theta, r)\, \mathrm{d}r \tag{10.8}$$

$$g_i = \frac{1}{\sigma_{s,i}} \int_{r_{\min}}^{r_{\max}} C_{s,i}(r) g_{p,i}(r)\, \mathrm{d}r \ . \tag{10.9}$$

Once the scattering properties have been determined for each individual particle inclusion, the bulk properties are computed using a weighted average [Grenfell 1983; Light et al. 2004]:

$$p(\theta) = \frac{1}{\sigma_s} \sum_{i \in A} \sigma_{s,i} p_i(\theta) \tag{10.10}$$

$$g = \frac{1}{\sigma_s} \sum_{i \in A} \sigma_{s,i} g_i \ . \tag{10.11}$$

Considering the number of Lorenz-Mie expressions required to approximate the true Lorenz-Mie phase function $p(\theta)$, it is only practical to either tabulate the

phase function or use a mean number density for each particle inclusion. It is also possible to use the ensemble asymmetry parameter g in Equation 10.11 with one of the standard phase functions, e.g. the Henyey-Greenstein phase function which is defined by [Henyey and Greenstein 1940]

$$p(\theta) = \frac{1}{4\pi} \frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{3/2}} \quad .$$

A more exact option is to use a multi-lobed phase function where the Henyey-Greenstein function replaces $p_i(\theta)$ in Equation 10.10.

We have now found formulae for computing all the bulk optical properties of materials that are composed of homogeneous spherical particles. Input for the formulae are the complex refractive indices of host medium and particles as well as size distributions of the particles. If measured indices of refraction are not available, we have seen in Chapter 8 that it is possible to compute complex indices of refraction using the chemical formula of a substance. The main job in computing the optical properties of a material is then to determine the types of particles that the material is composed of and their size distributions.

Chapter 11

Colour

"It's like this," he said. "When you go after honey with a balloon, the great thing is not to let the bees know you're coming. Now, if you have a green balloon, they might think you were only part of the tree, and not notice you, and if you have a blue balloon, they might think you were only part of the sky, an not notice you, and the question is: Which is most likely?"

A. A. Milne, from *Winnie-The-Pooh*

While it has, in general, not been stated explicitly, all the optical properties and the radiometric quantities that we have discussed have been wavelength dependent. The absorption coefficient represents an absorption spectrum, the scattering coefficient a scattering spectrum, etc. If we sample radiance values at different wavelengths in a rendering, we get a spectrum in each pixel. In the following we will shortly explain how spectra translates to the trichromatic colours often used in graphics (Sec. 11.1). Afterwards we draw some conclusions on this part of the thesis (Sec. 11.2).

11.1 Trichromatic Representations

As mentioned in Section 2.2, it was originally proposed by Young [1802], and later confirmed by Helmholtz [1867], that the eye has three receptors. It was conjectured that all the colours we are able to see are a combination of three principal colours: Red, green, and blue. While the real colour receptors in the

eyes do not precisely measure red, green, and blue, the Young-Helmholtz theory is essentially true. It is indeed possible to represent all visible colours by three values. The CIE (Commission Internationale de l'Éclairage) has chosen a set of standard conditions for measuring the human response to colour. Following these standard conditions, different sets of curves have been measured which translate a spectrum to a trichromatic representation.

Three curves have been made which translate a spectrum to a red, a green, and a blue colour. They are referred to as the CIE RGB colour matching functions. Described very concisely, the colour matching functions were determined experimentally using three almost monochromatic light sources: $r = 700\,\text{nm}$, $g = 546.1\,\text{nm}$, and $b = 435.8\,\text{nm}$. An observer was shown a target colour and then mixed the light from the three monochromatic sources in different intensities until the target colour was matched. To match all the colours, it turned out that it was sometimes necessary to add some light directly to the target (instead of in the mix of r, g, and b). This is interpreted as "negative" colour and it is the reason why the RGB colour matching functions are sometimes negative. To avoid negative colour values, the CIE also has three curves called the XYZ colour matching functions. These have been chosen such that they represent all the visible colours without being negative.

According to Stockman and Sharpe [2000], the "most secure and extensive of the available color matching data" is the 10° RGB colour matching functions of Stiles and Burche [1959]. The 10° mean that the size of the target colour is ten degrees around the center of the observer's visual field. A slightly corrected version of Stiles and Burche's RGB colour matching functions was included in the work of Stockman and Sharpe [2000]. These are the curves I have used in my implementations. Let us use the notation $\bar{r}(\lambda)$ for the red colour matching function, $\bar{g}(\lambda)$ for the green function, and $\bar{b}(\lambda)$ for the blue function. The way of computing RGB colour values from a spectrum is

$$\begin{aligned} R &= \int_{\mathscr{V}} C(\lambda)\bar{r}(\lambda)\,\mathrm{d}\lambda \\ G &= \int_{\mathscr{V}} C(\lambda)\bar{g}(\lambda)\,\mathrm{d}\lambda \\ B &= \int_{\mathscr{V}} C(\lambda)\bar{b}(\lambda)\,\mathrm{d}\lambda\ , \end{aligned}$$

where $\mathscr{V}$ denotes the interval of visible wavelengths (approximately from 380 nm to 780 nm) and $C(\lambda)$ is the spectrum that we want to translate to RGB.

Computer monitors do not emit light at the same almost monochromatic red, green, and blue colours as the ones used to find the colour matching functions. One way to show the difference between colour spaces is to plot them in a chro-

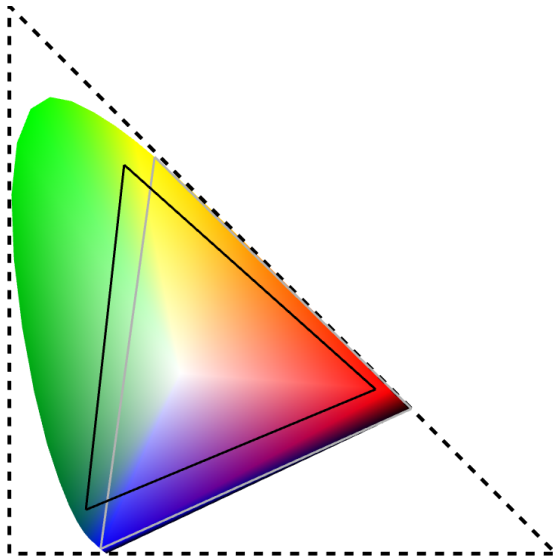


Figure 11.1: *The chromaticity diagram. The stippled triangle illustrates the XYZ colour space. The grey triangle illustrates the RGB colour space. The black triangle (which is not stippled) shows the gamut of a standard CRT display.*

maticity diagram. A chromaticity diagram has an arced curve which corresponds to all the monochromatic colours (see Figure 11.1). In-between the monochromatic colours, the diagram illustrates how spectra with a mix of monochromatic colours form the colours that we see. A trichromatic colour space is a triangle in the chromaticity diagram. The XYZ triangle encompass the entire chromaticity diagram. This means that some XYZ colours are not visible. The RGB triangle does not go outside the shape formed by the arc of monochromatic colours. This is not surprising since it was made with almost monochromatic light sources. Since the shape with all the visible colours is convex, any trichromatic colour space defined by points inside the shape will require negative colour values to capture all the visible colours. Figure 11.1 shows the difference between the trichromatic colour values given by the RGB colour matching functions and by the XYZ colour matching functions.

A CRT (Cathode Ray Tube) display is limited by the radiation capabilities of the phosphors that it has been build with. An LCD (Liquid Crystal Display) is limited by the radiation spectrum of the backlight it has been build with. In general a display cannot show colours outside the triangle spanned by the three primary colours that it uses. This triangle is called the *gamut* of the display. Figure 11.1 shows the gamut of a standard CRT display. The CRT gamut and the LCD gamut are fairly close to each other in most colour regions [Sharma 2002]. It should be noted that the shape and orientation of the typical display

gamut and the triangle defining the RGB colour space are quite similar in shape and orientation. This means that the RGB colour matching functions will, in most cases, translate a spectrum to RGB colour values that make sense when displayed on a computer screen. If one wishes to perfect the colours displayed on a screen, the solution is to make a 3×3 matrix which transforms XYZ or RGB colour values to the gamut of the monitor. This requires knowledge of the gamut and the monitor white point. How to find the transformation matrix, when this information about the display has been obtained, is described by many authors. See, for example, the recipe by Glassner [1995, Sec. 3.5].

If we choose to transform our trichromatic colour values to suit the gamut of a display, it is important that we do not transform them until we have the final pixel colours for an image. It is common practice in graphics to use trichromatic colour values throughout a rendering. This means that absorption spectra, scattering spectra, etc. are all translated to a trichromatic colour space. If we use this approach instead of spectral rendering, it is incorrect to us monitor specific colour values. The rendering should be done with CIE XYZ or CIE RGB colour values. Adaption of the colours to a specific display is the final step. If we use the RGB colour matching functions, for example, to translate an absorption spectrum to RGB, we may run into problems with negative values. Negative absorption does not make sense (in this context). Therefore the XYZ colour matching functions are the safest choice. The disadvantage of the XYZ colour space is that we have to translate the values to some other space before we display them. This makes the XYZ colour space less convenient. In the rendering programs made for this thesis, I have chosen to use either CIE RGB rendering or full spectral rendering with conversion from spectrum to RGB using the RGB colour matching functions. I have not run into negative colour values. If negative colour values are encountered, they should be corrected using a gamut mapping technique. Glassner [1995, Sec. 3.6] describes some of these techniques and gives pointers to other references.

11.2 Conclusions

At the very beginning of this part (Chapter 8), we asked about the level at which to use measurements. There is no eternally true answer to this question. I would say that measured properties should be used whenever they capture precisely the materials that we need. This is rarely the case for measurements at a macroscopic scale. The more microscopic the scale, the more often measurements are available that capture everything we need. To describe a small homogeneous particle, the particle shape and the index of refraction is everything we need. If the index of refraction has been measured for a substance, there is really no

need to compute it. If not, we found that it is possible to obtain an approximation using the chemical formula for the material. It requires some information about the polarisability of atoms, but the list of atoms is relatively short, so there is a good chance that this information is available in the literature. If we want to know how an index of refraction varies with the physical conditions of the medium, and if this dependency has not been measured, there may be a way of approximating it using information about the atoms in the material. The atoms can also tell us a lot about the emission and absorption properties of matter. It is, however, a subject which requires that we work with quantum electrodynamics.

The next step, where we have to choose between measurement and computation, is where we go to a macroscopic description of scattering. In Chapter 9 we described how to compute the scattering properties of a single particle. It is extremely difficult to measure the scattering of a single particle, so at this point we probably have to choose computation. In Chapter 10 we went from scattering by a single particle to scattering by a cloud of particles. Here we could choose measurements, but now we are at a level which is so macroscopic that we have to be lucky to find measurements of precisely the particle cloud that we are interested in. Sometimes we may find measurements for a few wavelengths or for a trichromatic colour space, but it is not likely that we will find measurements for an entire spectrum. An option is then to interpolate and extrapolate the measurements, but we could easily spend as much time doing that as the time it would take to compute the properties from a microscopic description. Therefore I think that it is a good idea to compute scattering properties. It gives many advantages. It enables us to make models that capture how the bulk optical properties change as we change the contents of a medium.

In this chapter, we have looked at trichromatic colour spaces. It is quite common in graphics to represent the properties of materials as RGB colour values. Techniques for measuring the scattering properties of materials in RGB have been presented by Jensen et al. [2001], Tong et al. [2005], and Narasimhan et al. [2006]. They all capture light scattered by materials using cameras. The RGB image data from the camera is used to obtain scattering properties in RGB format. This is an interesting low cost way of measuring scattering properties. The methods are, however, limited in different ways. The method by Jensen et al. [2001] and Tong et al. [2005] are only suitable for highly scattering fairly isotropic materials. They are based on diffusion and do not capture the asymmetry parameter. The method by Narasimhan et al. [2006] is based on dilution and therefore only works for fluids. This means that it is also worth computing optical properties even if we only need the properties in RGB format.

Measurement will always be difficult, especially if we also want to know how optical properties change with the physical conditions of the medium. In addi-

tion, measurement requires that we have a copy of the material that we want to render. This is not always obtainable. Computation, on the other hand, requires input that are not always available. The conclusion is that we will always need both measurement and computation of optical properties. They are complementary not competitors.

Part III

GEOMETRY

Chapter 12

Shapes

endless forms most beautiful and most wonderful have been, and are being evolved.

Charles Darwin, from *The Origin of Species*

Given the following input concerning a scattering material:

- Relative particle contents (volume fractions)
- Particle shape for each particle type (if not sphere)
- Size distribution for each particle type
- Complex index of refraction for each particle type
- Complex index of refraction for the host medium,

the theory described in Part II enables us to compute the following output:

- Bulk index of refraction
- Bulk extinction coefficient
- Bulk scattering coefficient
- Ensemble phase function

- Ensemble asymmetry parameter.

This output is all we need to render the scattering material using the theory described in Part I. Therefore, if we find the input, we have an appearance model. If rendering is all we want to do, we could skip straight from here to the results in Part IV. On the other hand, if rendering were all we wanted to do, we could perhaps use a simpler heuristic scheme for computing optical properties instead of the theory described in Part II. From my point of view, the computation of optical properties for rendering is important, but it is not the most interesting aspect of this thesis.

The first two parts of the thesis have described the connection between microscopic and macroscopic optical properties. In my opinion, the most interesting aspect in this connection is the opportunity to learn about the meaning of the appearance of materials. In other words, the opportunity to learn what the appearance means with respect to the contents of the material, with respect to the size distribution of the particles, with respect to the temperature of the material, etc. However, to learn about these relations, we need a technique which allows us to extract the relation between one or two outputs and one or two inputs. The mapping from all the inputs to all the outputs is not enough. If we want to be able to analyse appearance, it is particularly important that we are also able to find the consequences with respect to the input when we vary the output. In this part of the thesis, we will build a framework for constructing versatile appearance models that are more than just a direct mapping from input to output.

What we are looking for is a model that does not distinguish between input and output. A model in which we can freely vary all parameters, and extract the relation between a subset of the parameters. This is exactly what we are able to do if we find a geometrical way to represent our appearance models. In the remainder of this chapter, we will, briefly, discuss geometry as it is used in graphics. Then we will introduce the concept of multidimensional shapes for representation of appearance models, and we will investigate how we can use this shape representation to make versatile appearance models.

Rendering has been discussed at various points throughout this thesis. We have discussed how to model light (Part I); how to model matter (Part II); but not how the geometry of a scene would be represented. We have referred to ray tracing as an approach which can capture the scattering by particles of arbitrary geometry (Sections 5.4 and 9.3), but we have not talked about how it is possible. Anyone who has worked in graphics will not be surprised that it is possible. The general approach is to use a large number of small triangles. If we let some triangles have common edges with others, we obtain a *triangle mesh*.

If there are no holes in the mesh, it represents the surface of a volume which could be a homogeneous particle of arbitrary shape.

The triangle mesh poses some problems. Unless we want to make objects with sharp edges, for example crystalline particles, we need an awfully large number of triangles to make an object look smooth. Even more triangles are needed if we want smooth reflections and refractions as well. This problem is usually solved by computing a normal for each vertex in the triangle mesh. Such a *vertex normal* is an average of the normals associated with the neighbouring triangle faces. When we want to find the *surface normal* where a ray intersects a triangle, we interpolate the vertex normals across the triangle using trilinear interpolation. In this way, we have a number of points (the vertices in the triangle mesh) which make out a discrete representation of a smooth surface.

Sometimes we need more than just a surface. If we want to model a heterogeneous medium, we need a volume representation. One option is to use a grid with a set of material properties for each grid cell. We then interpolate the material properties between grid cell centres to get a smooth volume, or another option is to smooth the transition between values in different grid cells using a filter kernel. There are many different grids that we can choose. The most obvious choice is a rectangular grid, but we could, for example, also choose a tetrahedron grid.

Suppose we want the material to change over time, then we need yet another dimension. An option is to use a grid like in three dimensions. We could choose a hypercube grid or a pentatope grid. A pentatope is the simplest primitive in four dimensions. It has five vertices just like the triangle has three and the tetrahedron has four. The simplest primitive is sometimes called a *simplex*, and it has $n + 1$ vertices in n dimensions. A cube has 2^n vertices in n dimensions. Thus we see that in any (finite) number of dimensions, we can make a discrete representation of a smooth shape.

What is a shape? Intuitively, it is a drawing in two dimensions or a physical object in three dimensions. Mathematically, we may choose to let an n-dimensional shape denote a collection of points that satisfy constraints involving n variables. Here we let a *constraint* denote a mapping between a number of arbitrary sets and the set of Boolean values (true and false). The variables given as arguments to the mapping are referred to as the *involved* variables. It does not have to be spatial variables. A shape can represent a relation between any set of variables. It might be a relation between Boolean variables, or a relation between the optical properties of a medium. Every point in n-dimensional space is a configuration which is either valid or invalid under the given constraints. The set of valid configurations is a shape. A sphere is a simple shape which consist of all the points in (n-dimensional) Euclidian space for which the distance from

a centre point is less than a given radius. The constraint which makes a sphere is a simple equation, but many shapes are not as easily specified. We may need many constraints to describe a shape. Doing arbitrary shapes, we will most often not be able to represent them by simple mathematical expressions.

Why are more than three or four dimensions necessary? Because all the properties of an object are interrelated. The microscopic structure of a material governs the properties of the material. The material properties sometimes change the macroscopic physical shape. The physical conditions, such as temperature and pressure, govern the microscopic structure as well as the material properties, etc. If we are able to describe these relations as a multidimensional shape, we have an incredibly powerful appearance model. An appearance model which does not distinguish between input or output.

It poses a number of challenges to describe relations between the properties of an object as a constraint problem in many dimensions. First we need to construct the multidimensional shape. Secondly, we need to compress and store it in one way or another. Thirdly, we need to be able to draw conclusions based on the final shape, that is, we want to extract information about the object when it is placed under different circumstances. In Chapter 13 we will discuss a way in which we can represent the multidimensional shapes. There are many different ways to represent them, but I find an array-based representation particularly instructive and simple to work with. In Chapter 14 we will discuss general geometric operations that make us able to perform the three challenging tasks just described.

It should be noticed how the multidimensional shapes are analogous to particle systems in quantum mechanics. In Chapter 3 we saw how the properties of all the particles in a system are combined into a many-fold infinite-dimensional problem. These systems were handled using operators. Analogously we will use operators to handle our shapes. Like the creation and annihilation operators, we will formulate operators for creating shapes and for reducing shapes. The most common reduction (or extraction) will be to pick a subspace containing the material properties or physical shape under some specific conditions. Essentially it is the same as fixing some of the arguments in a function describing an object. Alternatively we might also want to know the general relation between a small subset of the involved variables. This requires a large amount of work if the shape is described by a complicated mathematical expression. In a geometrical representation such conclusions are drawn by a simple orthogonal projection. In summary, the geometrical approach offers an adaptable and versatile way of representing the relation between physical conditions, material properties, and appearance of an arbitrarily shaped object.

Chapter 13

Boolean-Valued Arrays

By dimension we understand nothing other than the mode and reason according to which some subject is considered to be measurable; so that not only length, width, and depth are the dimensions of body, but also gravity is a dimension according to which subjects are weighed, speed is a dimension of motion and so on indefinitely.

René Descartes (1596 – 1650), from *Rules for the Direction of the Mind*

As mentioned in the previous chapter, we let a shape denote a set of points in n-dimensional space. The points satisfy constraints involving n variables. It is our quest, in this chapter, to find a feasible way of representing shapes. A way which is close to the *geometrical image* of the shape and yet compact. By geometrical image, I mean the shape when plotted in the coordinate system spanned by the involved variables. In general, it is not feasible to represent the shape by the geometrical image itself. If we work with continuous domains, there are infinitely many points in the shape. It is, however, illustrative to learn about the geometrical image by considering simple discrete domains. In the following we will try to build some mathematical structure around the geometrical images of simple discrete domains.

A point that belongs to a shape is a value of truth in the n-dimensional coordinate system which is spanned by the involved variables. All the points that do not belong to the shape are values of falsehood in the same coordinate system. Suppose we let $A_1, A_2, \ldots$ denote arbitrary sets. Then an n-dimensional shape

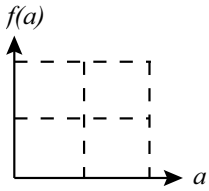


Figure 13.1: *The discrete coordinate system in which we draw the geometrical image of Boolean functions* $f : \{\mathrm{o},\mathrm{l}\} \to \{\mathrm{o},\mathrm{l}\}$ *of a single argument* a*.*

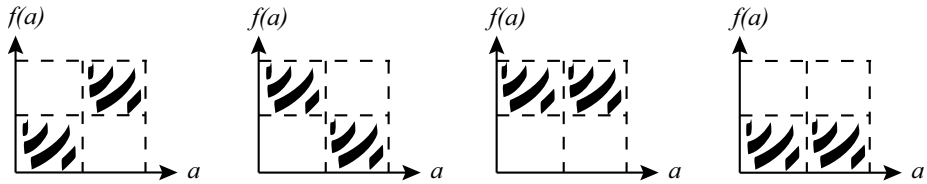


Figure 13.2: *All possible functions* $f : \{\mathrm{o},\mathrm{l}\} \to \{\mathrm{o},\mathrm{l}\}$*.*

is also a function

$$f : A_1 \times \cdots \times A_n \to \{\mathrm{o},\mathrm{l}\}$$

from n arbitrary sets into the set of Boolean values. Functions of this type are referred to as *Boolean-valued functions*. To distinguish integers from Boolean values, the letters l and o are used to denote the Boolean values true and false.

To start out as simple as possible, let us work with shapes that describe relations between Boolean variables. A Boolean-valued function $f : \{\mathrm{o},\mathrm{l}\}^n \to \{\mathrm{o},\mathrm{l}\}$, which takes only Boolean values as arguments, is simply called a *Boolean function*. To illustrate the small conceptual difference between discrete and continuous domains, let us have a look at simple coordinate systems describing Boolean functions. The coordinate system describing a Boolean function with a single argument is shown in Figure 13.1. It is easy to draw all possible functions f in this discrete coordinate system. They are shown in Figure 13.2. To draw a shape involving more than one variable, we add more axes to the coordinate system. As the number of variables increases beyond three, it becomes almost impossible to imagine what the shape looks like. Therefore we need something more abstract than a mere drawing to describe multidimensional shapes.

One way to write the geometrical images of the functions in Figure 13.2 is using *arrays*. Trenchard More [1973a; 1973b; 1973c] founded a new branch of discrete mathematics with the nested rectangular array as the primary notion. His theory is called *array theory* and it provides a rigorous mathematical foundation for formulating arrays and operations on arrays. Like a coordinate system, an

Geometrical image	State	Propositional function
o l	Truth	Affirmation
l o	Falsehood	Negation
l l	Indefinite	Tautology
o o	Impossible	Contradiction

Table 13.1: *An interpretation of the four Boolean functions in Figure 13.2.*

array is a rectangular arrangement. Thus a one-dimensional array has a single axis. An n-dimensional array has n axes. The operations of array theory are defined to work for arrays with an arbitrary finite number of axes. This gives a great advantage when we want to define operations that work in general on shapes of an arbitrary number of dimensions. Another advantage of an array-theoretic description of shapes is the tight connection of the array representation to the geometry of the shape. And, in addition, arrays are suitable structures for storage and arrangement of data on a computer.

The array representations of the functions in Figure 13.2 are the pairs of Boolean values (o l), (l o), (l l), and (o o) respectively. The structure of the array maps the argument to the function value. We have

$$0 \text{ pick } (a\ b) = a \quad \text{and} \quad 1 \text{ pick } (a\ b) = b \ , \tag{13.1}$$

where $a, b \in \{\text{o}, \text{l}\}$ are arbitrary Boolean values and the binary operation I **pick** A returns the *item* of the array A at the index I. Suppose A_f is the array representing a function $f : C \to D$ from the set C into the set D. Let us define a bijection $I_f : C \to C_I$ from the domain C of the function f onto the set of indices C_I into the array A_f by

$$I_f(x) = I \quad , \quad x \in C \quad , \quad I \in C_I \ .$$

Then the operation **pick** corresponds to giving an argument x to a function f:

$$I_f(x) \text{ pick } A_f = f(x) \ .$$

In the following, we will refer to the function I_f as the *index transform* for the domain of the function f.

We can state the name of the variable which an axis represents explicitly. This is done using the following notation:

$$\overline{\text{o l}}^{\,P} \qquad \overline{\text{l o}}^{\,P} \qquad \overline{\text{l l}}^{\,P} \qquad \overline{\text{o o}}^{\,P} \ , \tag{13.2}$$

where P is the variable.

To make these simple shapes a little less abstract, let us assign some meaning to them. Since they are Boolean functions, we can think of them as representing propositional functions. Then they describe a relation involving a single propositional variable. As an example, consider the proposition

$$P: \quad \text{I am flying} \ .$$

What would be the meaning of the four functions in Figure 13.2 if P were the argument? Suppose the value returned by the function concerns the truth or falsehood of the proposition, then insertion of the different possible values of P has the following results. The first function (o l) says that if I am not flying, I am not flying; if I am flying, I am flying. This is an *affirmation* of the proposition. The second function (l o) says that if I am not flying, I am flying; if I am flying, I am not flying. This is a *negation* of the proposition. The third function (l l) says that either way I am flying. This is termed a *tautology*. The fourth function (o o) says that I am not flying no matter what. This is a *contradiction*. In the world of propositional functions, we now know the meaning of these four primary functions. If we think of the functions as shapes which describe the valid states in a configuration space, the terms attached to the four functions are different. Table 13.1 summarises the two different interpretation of the four shapes.

Having an interpretation of the discrete shapes, makes it more interesting to go through the steps leading to a representation which allows continuous domains. In the following section we will increase the number of dimensions.

13.1 Arrays

If we extend the notation of the previous section, we can define a few shapes involving two Boolean variables P and Q by

$$\begin{array}{|cc|}\hline o & o \\ o & l \\ \hline\end{array}^{Q}_{P} \qquad \begin{array}{|cc|}\hline o & l \\ l & l \\ \hline\end{array}^{Q}_{P} \qquad \begin{array}{|cc|}\hline l & l \\ o & l \\ \hline\end{array}^{Q}_{P} \qquad \begin{array}{|cc|}\hline l & o \\ o & l \\ \hline\end{array}^{Q}_{P} \ ,$$

where we have chosen to display the first axis vertically and the second axis horizontally. The reason for this choice will be explained shortly. First let us see what these shapes mean. If we think of the arrays as representing Boolean functions, and use **pick** and the index transform with the first of the arrays,

say, we get the truth table for the relation between P and Q that it describes:

P	Q	result
o	o	o
o	l	o
l	o	o
l	l	l

Interpreting P and Q as propositions, it is immediately obvious that this first of the four shapes describes the propositional formula $P \wedge Q$. The second shape describes $P \vee Q$, the third $P \Rightarrow Q$, and the fourth $P \Leftrightarrow Q$.

The next step is to deal with shapes that involve three variables, and eventually n variables. So we have to find a general rule for displaying the array representing an n-dimensional shape. More [1981] proposed that a multidimensional array is written with axes alternately vertical and horizontal starting with the last axis horizontal and innermost. Thus we see the reason why the second axis was displayed horizontally when we considered a two-dimensional array. To learn the notation, let us have a look at an array representing a shape which involves the variables P, Q, and R:

$$\begin{array}{c|cc|cc} & & & & P \\ & & R & & R \\ Q & \text{l} & \text{l} & \text{o} & \text{o} \\ Q & \text{o} & \text{l} & \text{o} & \text{l} \end{array} \quad . \tag{13.3}$$

The meaning of this shape is not immediately obvious, nor is it obvious where it came from. Did I choose the points that define the shape (the truth values) at random? Actually, I did not, but the point is that shapes are not interesting if we do not assign a meaning to them. The shape itself is not interesting, the relations that it implies between the involved variables are interesting. So there are two objectives to find general operations for: (1) Conversion of a set of constraints into the array representation of a shape and (2) derivation of relations that a shape implies between the involved variables (both in general and under specific conditions, where some variables are bound to certain values). We will look at the first subject shortly and at both subjects in Chapter 14.

The principle of well-ordering is fundamental in array theory [More 1979], and it seems to be fundamental in the human way of understanding the world. As More [1981] points out, this was also recognised by Abraham Fraenkel, one of the prominent developers of set theory. Fraenkel [1953, p. 172] wrote

> From a psychological viewpoint, there can be no doubt that somehow the ordered set is the primary notion, yielding the plain notion of set or aggregate by an act of abstraction, as though one jumbled together the elements which originally appear in a definite succession. As a matter

> of fact, our senses offer the various objects or ideas in a certain spatial order or temporal succession. When we want to represent the elements of an originally non-ordered set, say the inhabitants of Washington D.C., by script or language, it cannot be done but in a definite order.

Thus every set either has a natural ordering, or we can assign one to it. All physical measures certainly have a natural ordering. The notion of arrays, then, provides us with a convenient way (in comparison to set theory) of representing the natural ordering of measurable quantities. Even for more abstract quantities: "Arrays reflect one's geometric and pictorial perception of collections much more closely than sets" [More 1981, p. 2].

While array theory was originally developed as a theory for giving rigorous mathematical definitions of programming operations [More 1973a], it was also founded in the belief that the manifestation of data originates in physical objects [More 1979]. This connection between data and physical (or geometrical) objects was also recognised by Ole Immanuel Franksen [1984a] who, in his description of data as geometrical objects, connected the fundamental principles of physical and geometrical theories to More's array theory [Franksen 1984b]. These fundamental principles are the expansion and reduction of dimensionality to analyse the properties of geometrical shapes in the broadest sense of the word. The operations which implement the fundamental principles are the outer product (for expansion), the repeated-index transposition, and the inner product (for reduction). As examples, Franksen defines the gradient in terms of an outer transform, and the divergence in terms of an inner transform [Franksen 1984c]. Here the term *outer transform* is used to say that the structural operation is the same as an outer product, but the operation is allowed to be different from the conventional multiplication. Likewise the *inner transform* may be used with operations that are different from the conventional addition and multiplication. Another example, which also illustrates the application of the repeated-index transposition, is the representation of logic as invariant theory [Franksen 1979; Franksen 1984d]. The repeated-index transposition corresponds to the operation of picking of a diagonal hyperplane in a multidimensional array. We will return to the significance of picking diagonals later. Let us first investigate how the outer transform is useful in the first of the objectives described previously.

A cartesian product gives all the coordinates in a coordinate system. So the coordinate system is a way of ordering the cartesian product (Figure 13.1 is a simple example). Like the coordinate system, the array is also ordered. To convert a constraint into an array, we need an ordered cartesian product for arrays. The ordered cartesian product is called **cart** in array theory [More 1975; More 1976]. Instead of giving sets as arguments (as for the conventional cartesian product), the arguments given to **cart** are arrays containing the values that a

variable can attain. This means that there must be some ordering imposed on the values if they do not have a natural ordering. In the terminology advocated by Franksen [1984a], we refer to the array which holds all the values that a variable can attain as the *scale* of the variable. The following is an example which demonstrates an ordered cartesian product of the scales of three Boolean variables:

$$\text{cart} \quad \overline{\text{o l}}^{P} \quad \overline{\text{o l}}^{Q} \quad \overline{\text{o l}}^{R}$$

$$= \quad \overline{\begin{array}{|c|c|}\hline \text{o o o} & \text{o o l} \\ \hline \text{o l o} & \text{o l l} \\ \hline \end{array}}^{R}_{Q} \quad \overline{\begin{array}{|c|c|}\hline \text{l o o} & \text{l o l} \\ \hline \text{l l o} & \text{l l l} \\ \hline \end{array}}^{R}_{Q}{}^{P} \quad . \tag{13.4}$$

The tuple at each position in the array is also an array. These *nested arrays* are written in box notation, and they do not have a particular variable associated with them. The nested arrays are the coordinates of positions in the coordinate system that the array corresponds to. Contrary to the arrays which represent shapes, it should be noted that the arrays which result from an ordered cartesian product are not necessarily Boolean-valued.

A constraint maps a number of values to a Boolean value. Thus a constraint gives rise to a shape. If we want the array to describe a shape involving P, Q, and R, all we need to do is to invoke the constraint on each coordinate given by the cartesian product (13.4). Consider the following constraint:

$$C: \quad (P \Rightarrow Q) \wedge (Q \Rightarrow R) \ .$$

We can think of it as a function $f_C : \{\text{o},\text{l}\}^3 \to \{\text{o},\text{l}\}$. Conventionally we write $f_C(P, Q, R)$ to invoke the function on the set of arguments $P, Q, R \in \{\text{o},\text{l}\}$. The corresponding array-theoretic operation is

$$f_C \ \ (P \ Q \ R) \ .$$

This is the operation that we would like to invoke on each coordinate given by the cartesian product (13.4). There is an operator to perform this task which we will return to shortly. First note that all array-theoretic operations in principle are unary in nature. They always take an array as argument. Even the binary operation **pick** is unary because we have the following general rule in array theory

$$A \ f \ B \ C \ \cdots = f \ \begin{array}{|c|c|}\hline A & B \ C \ \ldots \\ \hline \end{array} \ .$$

This construction does not prevent us from talking about binary operations or n-ary operations. It just means that the operation expects the input array to hold the arguments as n nested items. Therefore we will still refer to an operation as being unary, binary, or n-ary, and when we refer to its arguments, we are referring to the arguments nested in the input array.

The fundamental operator which applies a function to each item of an array is called EACH [More 1975; More 1976]. It gives us a way of invoking the constraint C on each coordinate given by the cartesian product (13.4). Letting juxtaposition of two operations denote the composition of two operations, we have

$$\mathrm{EACH}(f_C)\ \mathrm{cart}\quad \overline{\mathrm{o\ l}}^{\,P}\quad \overline{\mathrm{o\ l}}^{\,Q}\quad \overline{\mathrm{o\ l}}^{\,R} \;=\; \overline{\begin{array}{cc} {\scriptstyle Q}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{R} & {\scriptstyle Q}\left|\begin{array}{c}\mathrm{o\ o}\\ \mathrm{o\ l}\end{array}\right.^{R} \end{array}}^{\,P} \;,$$

This is the shape (13.3) that we did not recognise the meaning of before. Now we know what constraint it describes. The procedure we have used to construct the array from the constraint is general. The ordered cartesian product works for any number of arbitrary arrays, whether they represent entire constraints or scales of variables. If we let f_n denote a Boolean-valued operation of $n > 1$ variables, then the corresponding shape is (in array representation)

$$\mathrm{EACH}(f_n)\ \mathrm{cart}\quad A_1\quad \ldots\quad A_n\ ,$$

where A_i, $i = 1, \ldots, n$, are arrays each with items of the same type as the type expected for argument i of the function f_n. Apart from this restriction regarding the type of the items, the arrays are arbitrary. This is important, because it means that we are allowed to build the shape piece by piece. To give an example, let us construct the shape of the constraint C again, but this time using smaller pieces.

We already know the shape of $P \Rightarrow Q$, nevertheless let us construct it again from the basics. We have

$$\mathrm{EACH}(\Rightarrow)\ \mathrm{cart}\quad (\mathrm{o\ l})\quad (\mathrm{o\ l}) \;=\; \begin{array}{cc} (\mathrm{o} \Rightarrow \mathrm{o}) & (\mathrm{o} \Rightarrow \mathrm{l}) \\ (\mathrm{l} \Rightarrow \mathrm{o}) & (\mathrm{l} \Rightarrow \mathrm{l}) \end{array} \;=\; \begin{array}{cc} \mathrm{l} & \mathrm{l} \\ \mathrm{o} & \mathrm{l} \end{array} \;.$$

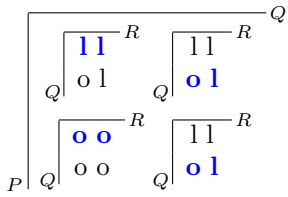
This is the shape of implication no matter what the name of the arguments are. Note that the operation is closely related to an outer product, except for the fact that we are using implication instead of multiplication. The operation is, in other words, an outer transform of implication. The general definition of the outer transform is [More 1976]

$$\mathrm{OUTER}(f) = \mathrm{EACH}(f)\ \mathrm{cart}\ ,$$

where f is an arbitrary operation. The operation $\mathrm{OUTER}(f)$ takes the same number of arguments as f. With this nomenclature, we construct the shape that describes the constraint C using two implications in the following way:

$$\mathrm{OUTER}(\wedge)\quad {\scriptstyle P}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{Q}\quad {\scriptstyle Q}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{R} \;=\; \overline{\begin{array}{cc} {\scriptstyle Q}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{R} & {\scriptstyle Q}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{R} \\ {\scriptstyle Q}\left|\begin{array}{c}\mathrm{o\ o}\\ \mathrm{o\ o}\end{array}\right.^{R} & {\scriptstyle Q}\left|\begin{array}{c}\mathrm{l\ l}\\ \mathrm{o\ l}\end{array}\right.^{R} \end{array}}^{\,Q} \;. \qquad (13.5)$$

It does not look like the shape we found before (13.3). The reason is, of course, that we have two axes representing the same variable. We will only ever use the positions in this array where the coordinates of the two Q-axes are the same. If we pick the diagonal between these two axes, we get the same shape as before. The diagonal is pointed out in the following illustration:



We will return to the operation that picks a diagonal in Section 14.2.

Although array theory is a discrete mathematical theory, it provides a very convenient way of thinking about multidimensional geometrical shapes. The operators and operations of array theory are abstractions that make it easier to get an understanding of general geometrical principles. Therefore we would like to extend the operators and operations of array theory to continuous scales, and eventually find a practical discrete representation. Extension of array theoretical concepts to shapes on continuous domains is the subject of the following section.

13.2 Continuous Domains

The moment we introduce continuous domains for the shapes, it is no longer easy to write up the array representation (as there will be infinitely many Boolean numbers). Therefore we will, to a larger extent, use the more abstract operators introduced in the previous section. It does still make sense to draw the shape in a coordinate system, but this does not really work for more than three dimensions. Thus, eventually, we have to work with the shapes in an abstract way.

As an example of a continuous shape, we will look at the circle (including interior). The constraint is well-known, it is the inequality:

$$\text{circle} : (x - x_0)^2 + (y - y_0)^2 \leq r^2 \ , \tag{13.6}$$

where $x, y, x_0, y_0 \in \mathbb{R}$ and $r \in \mathbb{R}_+$. In principle this is a five-dimensional shape describing all possible circles in two-dimensional space. We will limit our treatment to variables on intervals with end points that do not go to infinity. To do this in our example, we choose $x, y, x_0, y_0 \in [a, b]$ and $r \in [0, b]$. Of course we could have chosen distinct intervals for all the variables, but since they all refer

to the same Euclidian space, it makes sense to use the same end points for the intervals.

The easy way to construct the array representation of the circle shape is to use the Boolean-valued function corresponding to the circle inequality (13.6). If we call it $f_{\text{circle}}(x, x_0, y, y_0, r)$, we get

$$\text{OUTER}(f_{\text{circle}}) \quad \overline{a \ldots b}^{\,x} \quad \overline{a \ldots b}^{\,x_0} \quad \overline{a \ldots b}^{\,y} \quad \overline{a \ldots b}^{\,y_0} \quad \overline{0 \ldots b}^{\,r} \ .$$

This is not particularly interesting because we use the Boolean-valued function f_{circle} which describes the shape to construct the array representation of the shape. Instead, we should try to construct the circle shape piece by piece.

The circle inequality (13.6) clearly has three pieces which are all squared. The squared radius r^2 on the right-hand side is the simpler piece, so we will start with that. If we use the general method to find an array representation of r^2, we take an outer product of two interval arrays

$$\text{OUTER}(\cdot) \quad \overline{0 \ldots b}^{\,r} \quad \overline{0 \ldots b}^{\,r} \ ,$$

and pick the diagonal in the resulting array as both axes will represent the same variable. This is exactly the same as multiplying each item in the first array by the item at the same position in the second array. To avoid expanding an outer product of two arrays just to take a diagonal immediately after, it is convenient to have item-to-item operations. In order to construct an operator which applies an operation item-to-item, we use the operation called **pack** [More 1975; More 1976]. The operation **pack** uses the concept of nested arrays to pair the items at the same positions in a number of arrays. Using the same example as we did for **cart**, we have

$$\text{pack} \quad \overline{\text{o l}}^{\,P} \quad \overline{\text{o l}}^{\,Q} \quad \overline{\text{o l}}^{\,R} \quad = \quad \overline{\boxed{\text{o o o}\,|\,\text{l l l}}}^{\,P,Q,R} \ .$$

Note that when we use the general operation **pack** in this way, we assume that all three arrays represent the same variable. With **pack**, we define an operator similar to OUTER, but applying an operation item-to-item:

$$\text{EACHALL}(f) = \text{EACH}(f)\ \text{pack} \ ,$$

where f is an arbitrary operation taking more than one argument. Thus we write the array representing r^2 as follows:

$$A_{r^2} = \text{EACHALL}(\cdot) \quad \overline{0 \ldots b}^{\,r} \quad \overline{0 \ldots b}^{\,r} \ .$$

The other pieces of the circle inequality (13.6) also involve the OUTER operator. We have

$$A_{x-x_0} = A_{y-y_0} = \text{OUTER}(-) \quad (a \ldots b) \quad (a \ldots b) \ ,$$

where we use arrays that are not specific about which variables the axes represent. This is practical because the arrays representing the two terms on the left-hand side of the circle inequality are the same. To square them, we again use EACHALL. This gives

$$\begin{aligned} A_{(x-x_0)^2} = A_{(y-y_0)^2} &= \text{EACHALL}(\cdot) \quad A_{x-x_0} \quad A_{x-x_0} \\ &= \text{EACHALL}(\cdot) \quad A_{y-y_0} \quad A_{y-y_0} \ . \end{aligned}$$

The combined array representation of the circle shape is then

$$\text{OUTER}(\leq) \quad \big(\text{OUTER}(+) \quad A_{(x-x_0)^2} \quad A_{(y-y_0)^2}\big) \quad A_{r^2} \ .$$

This simple example demonstrates that there is only a small conceptual difference between shapes involving Boolean variables and shapes involving continuous variables. The same general principle applies in both cases: *It is possible to construct any shape using the* EACH *and* OUTER *operators and the operation of picking a diagonal.*

The EACH operator is included in this formulation of the general principle because it enables us to handle unary operations by applying them to each item of an array. For example, we could decide that y_0 is always 2. Then the array representing $y - 2$ is simply

$$A_{y-2} = \text{EACH}\,(f_{y-2}) \quad \overline{a \ldots b}^{\,y} \ ,$$

where $f_{y-2}(y) = y - 2$. In fact, we could use the unary function $f_{(y-2)^2}(y) = (y-2)^2$ and write

$$A_{(y-2)^2} = \text{EACH}(f_{(y-2)^2}) \quad \overline{a \ldots b}^{\,y} \ .$$

This illustrates that, just as for the outer transform, there is always a choice: we have to choose at what level the functional constraints should be converted into arrays. Put differently, there are many different ways of taking a shape to pieces.

In this section, and the previous section, we have outlined the general principle for converting functional constraints into the array representation of shapes. It should be realised that these functional constraints need not be mathematical expressions, they may as well consist of programs outputting one value or another for different inputs. The principle is very useful because it gives us a general way of obtaining the geometrical image of an arbitrary multidimensional shape. The example in this section is a five-dimensional shape representing all attributes of a circle in the plane. A movie of a moving circle that changes size would merely be a slice of this shape. This exemplifies the versatility of shapes. It brings us closer to the vision of having highly versatile appearance

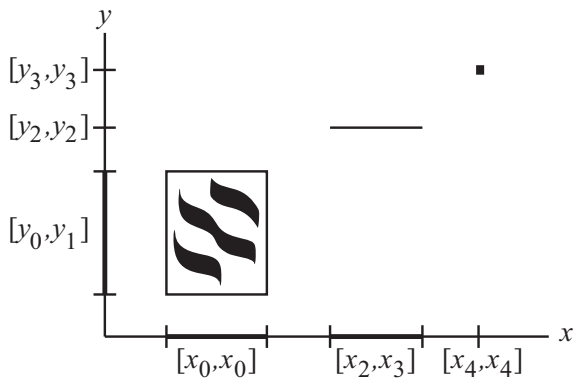


Figure 13.3: *The interval representation of continuous arrays.*

models. Unfortunately the array representation turns out to be infinite in size when we introduce continuous domains. This makes us unable to compute and store the complete arrays in a computer. In the same way that any other continuous entity has a discrete representation in a computer, we need a discrete representation for our continuous arrays. This is the subject of the following section.

13.3 Polynomial Representation

There are many ways of giving continuous entities a discrete representation. A large number of these representations have been explored in graphics. We use pixels to represent an image, triangles to represent a surface, voxel grids to represent volumes, etc. An ordinary rectangular grid is a very straightforward way of representing continuous arrays. Every grid cell has a value associated with it. These grid values represent the average value in a hypercubic part of the array. The resolution of the grid determines how accurately the grid represents the continuous array. The problem with this type of grid is that it grows quickly in size for every additional dimension in the shape. An adaptive grid would be more appropriate, but it is not immediately compatible with our operators. Another option is to represent only the values of truth (or only the values of falsehood). This corresponds to a mesh-based approach, where a mesh of hypervolumetric entities point out the values of truth (falsehood). A hypervolumetric entity could be a simplex in n-dimensional space.

An array-theoretic technique for representing continuous domains has been developed by Gert L. Møller [1995]. The idea is to represent continuous arrays

by collections of intervals. The intervals point out values of truth in the array projected on a given axis. Thus a shape which is a box in two dimensions, is represented by two intervals. The shape is the cartesian product of the two intervals. This gives a compact discrete representation of multidimensional shapes which are axis-aligned lines, points, boxes, or any combination thereof. See Figure 13.3. If we have a curve or an arced shape, we need infinitely many intervals. This means that we have to find a different discrete representation for smooth shapes.

The regular grid representation is convenient because all the operators (and operations) described in the previous section are exactly the same as for the original continuous array. Hence, the operators have not been described in vain. They give us an opportunity to work with approximate representations of multidimensional continuous arrays. Nevertheless, we will try to formulate a compact and more precise alternative to the grid representation using inspiration from the mathematical field of differential geometry.

Polynomial curves and surfaces comprise the traditional mathematical way of obtaining a discrete representation of smooth shapes. Using a discrete set of control points one obtains curves and surfaces that approximate any shape to any given precision [Gravesen 2002]. The problem is, of course, to find the right control points. A conventional approach is to compute a number of points in the shape (using the constraints) and then try to fit polynomial curves and surfaces to the points. This is difficult if we do not know the topology of the shape. How many curves and surfaces should we use for the fit? Instead, I propose that we use parametric hypersurfaces with one-dimensional control points to fill out a limited region of a multidimensional space with values. If we choose a convention such as *the surface is where the values are zero* or *the volume is where the values are less than zero*, we have an implicit representation of a multidimensional shape. The parametric hypersurface will be far more compact than a regular grid as it is described by a number of control points which corresponds to the polynomial degree of the hypersurface. It will also be much more precise if we find the right control points.

In our general principle for constructing arrays we start from the scales of the involved variables. For simplicity we assume that the scales of the variables are single, closed intervals $[a, b]$ (not necessarily with the same a and b for the different variables). If we think of a Bézier curve, this requires two control points: The end points of the interval. A Bézier curve is a parametrisation of the interval. The parameter $t \in [0, 1]$, which is given as argument to the Bézier curve, determines the position in the array (the index). The curve determines the value of the array at each position. See Figure 13.4. This gives a discrete representation of a one-dimensional continuous array. Let us call it a *polynomial*

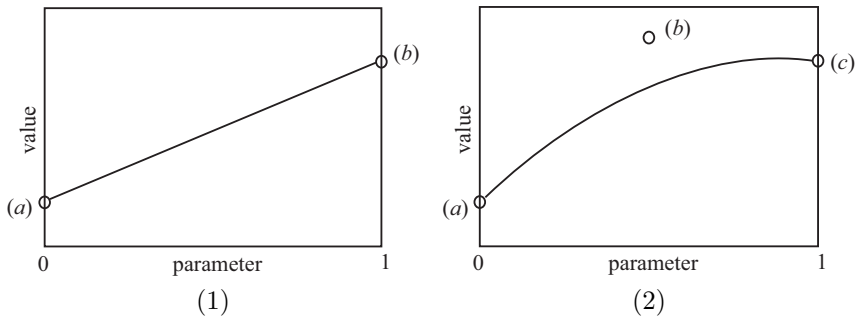


Figure 13.4: *An illustration of polynomial arrays. The parameter $t \in [0, 1]$ is the index into the array. The values of the array is determined by a Bézier curve. From left to right: (1) A one-dimensional polynomial array described by two control points (a) and (b). (2) A one-dimensional polynomial array described by three control points (a), (b), and (c).*

array and use the following notation:

$$(a \ldots b) = (\,(a)\ (b)\,),$$

where (a) and (b) are one-dimensional control points for a Bézier curve. As long as we have invoked no operations on the array, we only need two control points. If we start changing the array in ways that are more advanced than affine transformations, we need to introduce additional control points. The challenge in this representation is to find out when control points should be added and how the control points should be changed according to an arbitrary transformation of the array. This challenge is discussed in the following, but first we will show how a value in the continuous array is obtained from the control points of a parametric hypersurface.

To pick a value in the original continuous array using the polynomial representation, we use the *de Casteljau algorithm*. This is a very simple algorithm which has been described many times in the literature. The formulation of the algorithm by Gravesen [2002] is particularly suitable for translation into an array-theoretic operation. Gravesen defines two basic operators R and L. Taking a sequence of control points as argument, the operator R drops the last control point from the sequence, while the operator L drops the first control point. Array theory has two general operations called **front** and **rest** which drop the last and first item of an array, respectively. Thus it is simple to implement the de Casteljau algorithm in array theory. Following Gravesen, we define a forward difference operation:

$$\text{delta} = -[\text{rest}, \text{front}]\ ,$$

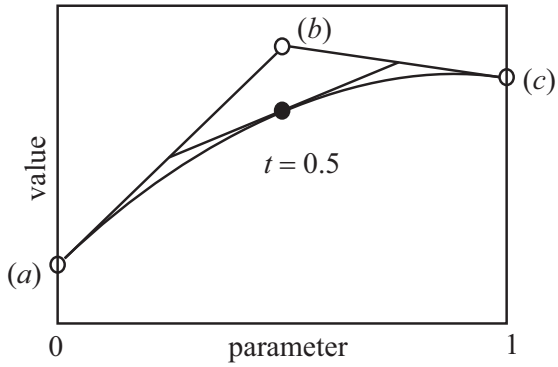


Figure 13.5: *Using some specific value for the parameter $t \in [0,1]$, the de Casteljau algorithm consists in successive applications of the de Casteljau operation. When there is only one value left in the array, it is the value at index t in the polynomial array. In this figure the polynomial array has three control points (a), (b), and (c), and $t = 0.5$.*

where the list notation $f\ [g,h]\ A$ constructs two arguments for a binary operation f by applying the operations g and h to the array A. In other words,

$$f\ [g,h]\ A = f\ \ (\,g(A)\ \ h(A)\,)\ .$$

The next operator defined by Gravesen is the de Casteljau operator $C(t)$. We define it as an array-theoretic operation by

$$t\ \text{opC}\ A = (\text{front}\ A) + (t \cdot (\text{delta}\ A))\ .$$

The de Casteljau algorithm is simply successive applications of the de Casteljau operation (using some specific value for the parameter $t \in [0,1]$). Figure 13.5 exemplifies how it works.

The de Casteljau operation applies to lists, that is, one-dimensional polynomial arrays (curves). To make the polynomial arrays useful in the context of multidimensional continuous shapes, let us generalise the de Casteljau algorithm such that it works for tensor product Bézier surfaces of n dimensions. First we choose the following convention. If we use the de Casteljau algorithm on an n-dimensional polynomial array parameterised by n variables $t_1, \ldots, t_n \in [0,1]$, we get an $(n-1)$-dimensional polynomial array parameterised by the variables $t_1, \ldots, t_{n-1} \in [0,1]$. In other words, the de Casteljau algorithm applies to the last axis of the array. To split an array into the part containing the front axes and the part containing the last axis, we use nesting. The fundamental operation in array theory for nesting is called **split**. The following is More's [1976] definition of **split**. The axes of an array are said to be on level 0, the axes of

the items of the array are on level 1, etc. The operation

$$\text{split } I \ A$$

divides level 0 of the array A into two levels. The indices I indicate which axes are to go down to level 1.

To point out the indices of the axes that we want to nest, we need a few extra operations. The indices of the axes of an array are obtained by the operation **axes**[1]. The two operations that complement **front** and **rest** are **last** and **first** which pick the last and first items of an array. Then the nested array with the last axis at the top level and all the remaining axes at level 1 is given by the operation

$$\text{polynest} = \text{split } [\text{front axes}, \text{pass}] \ ,$$

where the operation **pass** corresponds to identity (**pass** $A = A$). Then all we need to do is to apply the de Casteljau operation to the nested array.

One minus the length of an axis in a polynomial array is the *degree* of the corresponding dimension in the tensor product Bézier surface that the array represents. This degree denotes the number of times that we need to apply the de Casteljau operation. Array theory has an operation **shape** A which returns an array holding the length of each axis in A [More 1975].[2] Thus the degree of the nested array obtained using **polynest** is given by the operation

$$\text{degree} = -1 + \text{last shape} \ .$$

To apply an operation N times, we have the array-theoretic operator

$$N \text{ FOLD}(f) \ A \ , \tag{13.7}$$

where f is the operation to be applied N times to the argument A. This is all we need to define the general de Casteljau algorithm:

$$t \text{ deCasteljau } A = \text{first } ((\text{degree } A) \text{ FOLD}(t \text{ opC}) \text{ polynest } C) \ , \tag{13.8}$$

where we take the first item of the result to remove the extra level of nesting introduced by **polynest**. Note that t **opC** is the unary operation which results if we consistently choose t as the first argument of **opC**. This is referred to as *currying* t to **opC** [More 1993].

The definition (13.8) of the general de Casteljau algorithm demonstrates the ease with which array theory handles multidimensional problems. Generalisation

[1] The operation **axes** is not standard in array theory, but we can define it using standard operations as follows: **axes** = **tell valence**.

[2] The operation **shape** is not related to the concept of shape that we are discussing in this part of the thesis.

from one to two to n dimensions comes almost for free. With the **deCasteljau** algorithm we can define the equivalent of the **pick** operation for polynomial arrays. Let T hold the list of parameters $t_1, \ldots, t_n \in [0,1]$, and let A be a polynomial array, then we define

$$T \text{ polypick } A = \text{REDUCE(deCasteljau) append } T\ A\ ,$$

where the operation **append** T A places the array A as the last item of the array T and the operator REDUCE(f) B applies an arbitrary binary operation f to all items of the array B in right-to-left order. The reduction transform REDUCE has been defined formally by More [1979]. Reduction transforms of binary operations are well-known in other areas of mathematics. An example is the summation operator $\sum$ which is the reduction transform of the binary operation $+$.

Now that we know how to retrieve values from a polynomial array, let us investigate the general principles for construction of arrays to represent multidimensional shapes. To go from a one-dimensional to an n-dimensional polynomial array, we use the OUTER operator with addition ($+$), subtraction ($-$), and multiplication ($\cdot$). The polynomial array for $x + x_0$ is

$$\text{OUTER}(+) \quad \overline{(a)\ (b)}^{\,x} \quad \overline{(a)\ (b)}^{\,x_0} \quad = \quad {}_{x}\left|\begin{matrix}(a+a) & (a+b)\\ (b+a) & (b+b)\end{matrix}\right|^{x_0}\ ,$$

where $x, x_0 \in [a,b]$ and the contents of the array on the right-hand side are the control points for a tensor product Bézier surface. To be specific, let us set $a = -5$ and $b = 5$. The result is then

$${}_{x}\left|\begin{matrix}(-10) & (0)\\ (0) & (10)\end{matrix}\right|^{x_0}\ ,$$

which is simply a planar surface. This should not be surprising since we take the outer sum of two simple intervals. The multiplication case is more interesting. The polynomial array for x^2 is

$$\text{OUTER}(\cdot) \quad \overline{(-5)\ (5)}^{\,x} \quad \overline{(-5)\ (5)}^{\,x} \quad = \quad {}_{x}\left|\begin{matrix}(25) & (-25)\\ (-25) & (25)\end{matrix}\right|^{x}\ .$$

This is a surface which is shaped like a parabola. It is interesting because we cannot simply pick the control points along the diagonal to get the control points for the one-dimensional polynomial array representing x^2. The reason is that the degree of the Bézier surface has been increased. The array representing x^2 is of degree 2 and requires three control points. Inspecting the outer product array, we make the surprising observation that the opposite diagonal suggests the value of the control point that we should insert. Let us elaborate on this observation for a moment.

The outer product array is surely a valid way to find the polynomial array representing the product of two arbitrary polynomial arrays. The result, however, may contain axes representing the same variable. Using the de Casteljau algorithm, we obtain values in the continuous array that the polynomial array represents. But we always use the same parameter for the axes that represent the same variable. Thus we will have a much more compact representation if we are able to find the control points which represent the diagonal between the axes that represent the same variable. In the continuous array, the diagonal is simply the diagonal in the array, but in the polynomial representation we need to increase the degree by adding control points. The observation we have made leads to the assumption that the outer product array has all the information we need to find the control points that should be inserted when the degree of the polynomial array increases. Let us see if we can find a general way of getting the right control points from the outer product array.

For reasons that will soon be clear, we would like to work with the indices of the control points in the array. Therefore we introduce an operation which is called **grid**. It is a well-known operation in array theory (originally called **numerate** by More [1975]) which replaces all items in an array by their indices. To learn the notation, we find

$$\text{grid}\ {}_{x}\left\langle \begin{matrix} (25) & (-25) \\ (-25) & (25) \end{matrix} \right\rangle^{x} = {}_{x}\left\langle \begin{array}{cc|cc} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{array} \right\rangle^{x} .$$

To exploit our observation, we would like to apply an operation to the control points which are not in the diagonal. If we sum the indices in the grid, we get a set of indices that point out the items that are not in the diagonal:

$$\text{EACH}(+)\ \text{grid}\ {}_{x}\left\langle \begin{matrix} (25) & (-25) \\ (-25) & (25) \end{matrix} \right\rangle^{x} = {}_{x}\left\langle \begin{matrix} 0 & 1 \\ 1 & 2 \end{matrix} \right\rangle^{x} .$$

Let us assume that the indices 0, 1, and 2 in this array point out the control points for the one-dimensional polynomial array of x^2. To pick and place these control points in a new array, we introduce an operation called **polyfuseP** I A, where I holds the indices of the axes that represent the same variable. We will define the operation more precisely in Chapter 14. For now we will only look at the result, which is

$$\text{polyfuseP}\quad (0\ 1)\quad {}_{x}\left\langle \begin{matrix} (25) & (-25) \\ (-25) & (25) \end{matrix} \right\rangle^{x} = \left\langle (25)\ (-25)\ (25) \right\rangle^{x} .$$

Since there are only few control points, we have the opportunity to use the basis functions of the Bézier curves to verify that we really did find the polynomial

array for x^2. We have

$$\begin{aligned} & t \text{ polypick } (\,(25)\ (-25)\ (25)\,) \\ &= (1-t)^2 \cdot 25 + 2t(1-t) \cdot (-25) + t^2 \cdot 25 \\ &= 100t^2 - 100t + 25 \ , \end{aligned} \tag{13.9}$$

where the parameter $t \in [0,1]$ denotes the position in the array. The index transform is $t = (x+5)/10$. If we insert this in the equation (13.9), we get precisely what we hoped for

$$100\left(\frac{x+5}{10}\right)^2 - 100\frac{x+5}{10} + 25 = x^2 \ .$$

The procedure we have found also works if we want to take higher powers of an array. As an example, the polynomial array for x^3 is

$$A_r^3 = \text{OUTER}(\cdot)\ \overline{(-5)\ (5)}^{\,x}\ \left(\text{OUTER}(\cdot)\ \overline{(-5)\ (5)}^{\,x}\ \overline{(-5)\ (5)}^{\,x}\right)$$

$$= \overline{{}_x\left|\begin{matrix}(-125) & (125)\\ (125) & (-125)\end{matrix}\right|^{x}\ {}_x\left|\begin{matrix}(125) & (-125)\\ (-125) & (125)\end{matrix}\right|^{x}}^{\,x} \ .$$

The corresponding summed index array is:

$$\text{EACH}(+)\ \text{grid}\ A_r^3 = \overline{{}_x\left|\begin{matrix}0 & 1\\ 1 & 2\end{matrix}\right|^{x}\ {}_x\left|\begin{matrix}1 & 2\\ 2 & 3\end{matrix}\right|^{x}}^{\,x} \ .$$

Again the diagonal in the polynomial array gives the end control points and the remaining values in the array suggest the control points that we need to add to get the polynomial array for x^3. The result is

$$\text{polyfuseP}\quad (0\ 1\ 2)\quad A_x^3 = \overline{(-125)\ (125)\ (-125)\ (125)}^{\,x} \ .$$

It is easy to verify that this Bézier curve corresponds to x^3 as it should. This suggests that our new method for finding powers of polynomial arrays could be general. Assuming that it is, we can now find compact polynomial arrays that represent functions involving (integral) powers, sums and affine transformations.

Unfortunately the polynomial arrays cannot explicitly represent shapes and constraints (Boolean-valued functions). The problem is discontinuities where values jump from true to false (or oppositely). In principle, we would have to split the curve or hypersurface in two wherever there is a discontinuity. This is not very attractive. Instead, we choose a convention, as discussed previously, such that

we obtain an implicit (hyper)surface. It is not difficult to choose the convention. If we let Δ denote an equality or an inequality operation between polynomial arrays, we make the following replacement:

$$\text{OUTER}(\Delta) = \text{EACH}(0\ \text{CONVERSE}(\Delta))\ \text{OUTER}(-)\ ,$$

where the operator CONVERSE is a standard array theoretic operator which swaps the arguments of a binary operation [More 1981]. When we have found the polynomial array representing a shape, the EACH(0 CONVERSE(Δ)) is the convention. In Section 14.1, it will be shown that there are ways to work with the convention EACH(0 =) without actually finding the surface. Now we have enough information to construct, from the basics, a polynomial array representation of the circle shape.

First we find the polynomial array for $x - x_0$ and $y - y_0$:

$$A_{x-x_0} = A_{y-y_0} = \text{OUTER}(-)\ \ (\,(a)\ (b)\,)\ \ (\,(a)\ (b)\,) = \begin{matrix} (0) & (a-b) \\ (b-a) & (0) \end{matrix}\ .$$

If we square it using the discussed method[3], we get

$$A_{(x-x_0)^2} = A_{(y-y_0)^2} = \begin{matrix} (0) & (0) & ((b-a)^2) \\ (0) & (-(b-a)^2/2) & (0) \\ ((b-a)^2) & (0) & (0) \end{matrix}\ . \qquad (13.10)$$

The same method gives the polynomial array for r^2:

$$A_{r^2} = \text{polyfuseP}\quad (0\ 1)\quad \overset{r}{\left\lfloor \begin{matrix} (0) & (0) \\ (0) & (b) \end{matrix} \right.}\quad = \quad \overset{r}{\overline{(0)\ (0)\ (b)}}\ .$$

Putting it all together, we get the five-dimensional shape describing all circles in an arbitrary limited interval $[a, b] \times [a, b]$. The shape is

$$\text{EACH}(0 \geq)\,\big(\text{OUTER}(-)\quad \big(\text{OUTER}(+)\quad A_{(x-x_0)^2}\quad A_{(y-y_0)^2}\big)\quad A_{r^2}\big)\ ,$$

where the outer sum and difference are easily carried out. The resulting five-dimensional polynomial array holds $3^5 = 243$ control points. This means that it is a tensor product Bézier (hyper)surface of degree $2 \times 2 \times 2 \times 2 \times 2$. A value in the continuous array is obtained by the de Casteljau algorithm which uses one step for each degree of each dimension. This means that a value in the continuous array is obtained in $5 \cdot 2 = 10$ steps.

There are a some types of shapes that we have not considered. In particular, we have not considered shapes which involve a singularity (e.g. caused by division). Such shapes must be handled in a different way. Perhaps by a split of the

[3]This example is a little more complicated than the previous ones, because all the axes in the squared array do not represent the same variable. The more general method, which handles this case, is described in Section 14.2.

Bézier curve or surface at the singularity. Nevertheless, the idea of polynomial arrays is, in my opinion, inspiring. They combine the mathematical properties of tensor product Bézier surfaces and the structural properties of arrays. As we will see in the following chapter, the polynomial arrays enable a simple procedure for constructing multidimensional shapes which does not require algebraic manipulation of the constraints. It is a procedure that is readily automated and the discrete representation is compact. Since an array is also a tensor product Bézier surface, only few simple steps are required to query if a point is part of the continuous shape, or to obtain an implicit surface representation of the shape.

In this chapter we have seen that a Boolean-valued array is a way of representing a shape which is close to the geometry of the shape. In fact, the coordinates of the truth values in the Boolean-valued array are the points in space belonging to the shape. This means that we are able to obtain geometry or appearance, or whatever the shape describes, very quickly by simply picking a slice of the array. Since the array practically is the same as the geometry itself, it makes sense to handle the array in the same way as we handle geometry. Most importantly, it enables us to use the fundamental concept of projection. We will discuss the importance of projection and other geometric operations in the next chapter. The overall idea, in using geometric operations, is to derive the relation between a subset of the variables involved in a multidimensional shape. This is a difficult task if we only have separate constraints that have not been unified in a coordinate system or an array as geometry. When the geometry is available it is simply a matter of picking subspaces and doing projections.

Chapter 14

Geometric Operations

. . . perhaps the large is contained in the small,
life as a day, from night to night,
a drop is an ocean, a seed a world . . .
– then I understand myself and my life . . .

Jacob Martin Strid, from *Dimitri 9mm*

Let us briefly recall the quantum electrodynamics described in Chapter 3. To describe the state of a quantum particle, we need a probability amplitude for each possible combination of position (or definite momentum) and angular momentum of the particle. This means that a single photon has a two-fold infinite number of possible states (it has infinitely many possible positions and two possible angular momenta). If we think of all the possible states as a shape, just a single particle is incredibly difficult to describe. A system of particles is much worse. How is it possible to handle such complicated shapes? The answer was partly given in Chapter 3. We only look at one particular system state - one particular event - at the time. Even if we consider only one event at the time, we still have to take all the physical constraints of the system into account. This was accomplished in Chapter 3 using operators.

The creation and annihilation operators, $\hat{\psi}^\dagger$ and $\hat{\psi}$, enable us to keep track of a system in an abstract way. If we denote the vacuum state $|0\rangle$, we construct an N-particle system state by [Milonni 1994, formula generalised from examples]

$$|\Psi\rangle = \int \dots \int \phi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N; t) \hat{\psi}^\dagger(\boldsymbol{x}_1, t) \dots \hat{\psi}^\dagger(\boldsymbol{x}_N, t) |0\rangle \, \mathrm{d}\boldsymbol{x}_1 \dots \mathrm{d}\boldsymbol{x}_N \ ,$$

where ϕ denotes the probability amplitude for the particles to be at positions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ at time t. We can think of the creation operators in the formula as outer products of particle state vectors forming a multidimensional array which describes a specific system state $|\Psi\rangle$. The function ϕ determines the values in the array when the integration is carried out. From this point of view the annihilation operator $\hat{\psi}$ eliminates a part of the array corresponding to one particle by *projection* of this part of the array on the remaining axes in the array.

The Schrödinger equation (3.4) describes the constraints imposed on a system by the fundamental physical law of energy conservation. In the Schrödinger equation the Hamiltonian operator (3.25) is applied to a state vector to obtain the total energy of the system state. In Section 3.3 we found the Hamiltonian operator for the interaction between a photon field and a charge field. How does it work? It uses the annihilation operator to extract information about the energy of the system state. This illustrates that projection and outer product operators are fundamental in describing the incredibly complicated systems encountered in quantum electrodynamics.

In the previous chapter we saw that the array representation of shapes are created using an OUTER operator. This is quite analogous to the way of creating state vectors in quantum electrodynamics. The difference is that the array represents the entire system whereas the state vector $|\Psi\rangle$ only represents a specific system state. The state vector is only a specific system state because it stores only one complex probability amplitude for each particle state. It does not store all possible probability amplitudes. To represent the entire quantum mechanical system, we would need an additional axis of complex values for every value in the state vector. This means that we can think of the state vectors as complex-valued arrays instead of Boolean-valued arrays. Otherwise state vectors and arrays which represent shapes are similar. Since they are similar, it is not surprising that orthogonal projection is as fundamental a way of extracting information about a shape as it is a fundamental way of obtaining information about a system of quantum particles.

In this chapter we will describe geometric operations for handling multidimensional shapes. There are three fundamental types of operations: Outer sums, products, etc. for creation of arrays (Sec. 14.1), colligation for removal of redundancy in an array (Sec. 14.2), and projection as well as the picking of slices for extraction of information (Sec. 14.3).

14.1 Creation

The general principle for creating arrays which represent shapes has already been described in Chapter 13. The following is a brief recap. A shape is a set of points in n-dimensional space which satisfy a number of constraints. Every dimension of the shape represents a variable which is involved in one (or several) of the constraints. A constraint is a Boolean-valued function. There is logical conjunction between all the constraints, or, in other words, the points which define the shape must satisfy *all* the constraints. The general principle says that the array representation of the shape is given by the constraints if we replace all variables by their scales, that is, arrays holding all the values that the variable can attain, and if we replace all operations f by OUTER(f). This procedure may leave several axes that represent the same variable. An operation for removing this redundancy is discussed in Section 14.2.

The general principle is all we need to create the *expanded* array representation of any shape. The variables that the shape involves span a coordinate system. Since the array holds a Boolean value for every point in the coordinate system, we refer to it as the expanded array. For shapes that involve many variables or variables of a "large" scale (e.g. continuous variables), the expanded array representation is not practical. Møller [1995] demonstrated that there are isomorphic representations which are significantly more compact. Not only did he find more compact representations, he also found operations for creating the compact representation without needing to expand the array. This reduces the complexity of the operations for creating arrays as well as the memory required to store the array. However, as mentioned in Section 13.3, Møller did not find a discrete representation for all types of continuous arrays. Therefore I introduced the polynomial representation of continuous arrays. And it seems to be the case that also for polynomial arrays there are operators working directly with the compact representation without the need to expand the array. This is comforting because it is rather difficult to expand a continuous array (as it would require infinitely many items).

In section 13.3 we limited our discussion of the polynomial representation to a single constraint which only involved continuous variables. Let us see if we can make it more general. There is nothing preventing us from using integers in combination with continuous variables. Suppose we have a continuous variable x and an integer variable A of the following scales

$$\overline{a \ldots b}^{\,x} \quad , \quad \overline{1\ 2\ 3}^{\,A} \ .$$

The polynomial array representing the constraint $x = A$ would be

$$\mathrm{EACH}(0 =) \quad \mathrm{OUTER}(-) \quad \overline{(a)\ (b)}^{\,x} \quad \overline{1\ 2\ 3}^{\,A} \ .$$

There is no problem in taking the outer product as long as we are very careful in keeping track of the axis which has infinitely many entries and the axis which has only three entries. The polynomial array that results is

$$\text{EACH}(0=)\ C_1 \quad = \quad \text{EACH}(0=)\ {}_x\left|\begin{matrix}(a-1) & (a-2) & (a-3)\\ (b-1) & (b-2) & (b-3)\end{matrix}\right|^A \ .$$

This corresponds to three Bézier curves that have been translated differently.

If we have another constraint $x^2 = B$, where B is of the scale

$$\overline{\begin{matrix}-2 & -1\end{matrix}}^B \ ,$$

there is a very simple way of combining the two constraints. We know the polynomial array for x^2 (cf. Section 13.3). It is

$$\overline{\begin{matrix}(aa) & (ab) & (bb)\end{matrix}}^x \ .$$

So let us consider the total polynomial array for the second constraint:

$$\text{EACH}(0=)\ C_2 \quad = \quad \text{EACH}(0=)\ {}_x\left|\begin{matrix}(aa+2) & (aa+1)\\ (ab+2) & (ab+1)\\ (bb+2) & (bb+1)\end{matrix}\right|^B \ .$$

To find the shape that describes the two constraints

$$\text{EACH}(0=)\ C_1 \quad \text{and} \quad \text{EACH}(0=)\ C_2 \ ,$$

we need to find the array representing the logical conjunction. Since the convention is that all items of both C_1 and C_2 are set equal to zero, this is not so difficult. The following equation is a general way of taking the logical conjunction of two arrays for which all items are set equal to zero:

$$\begin{aligned}&\text{OUTER}(\wedge)\quad (\text{EACH}(0=)\ C_1)\quad (\text{EACH}(0=)\ C_2)\\ &\quad = \quad \text{EACH}(0=)\\ &\qquad \text{OUTER}(+)\quad (\text{EACHALL}(\cdot)\ C_1\ C_1)\quad (\text{EACHALL}(\cdot)\ C_2\ C_2)\ .\end{aligned}$$

One should verify the validity of this equation. Taking the square of all items in the arrays, leaves only positive values or zeros (as long as the items are not complex numbers). Requiring that every possible sum of these items is zero, is the same as requiring that all the original items are zero.

There is an even simpler equation that enables us to take the logical disjunction of two arrays for which all items must equal zero:

$$\begin{aligned}&\text{OUTER}(\vee)\quad (\text{EACH}(0=)\ C_1)\quad (\text{EACH}(0=)\ C_2)\\ &\quad = \quad \text{EACH}(0=)\quad \text{OUTER}(\cdot)\quad C_1\quad C_2\ .\end{aligned}$$

These two equations make us able to handle disjunction and conjunction of all polynomial arrays which represent constraints based on equality. It is important to be able to handle equality constraints. For many cases (perhaps all) it is possible to describe an inequality constraint by a projection of an equality constraint from a space with one additional dimension. The circle, for example, is the projection of a three-dimensional sphere. Thus we have quite a general procedure for creating the polynomial array representation of shapes. If we require a more general representation, there is always the option of sampling points in the continuous arrays and interpolating between them. This is a less precise approach, but it has the advantages of simplicity. In the next section we will look at removal of redundancy in arrays. In particular, we will describe the operation of picking a diagonal.

14.2 Colligation

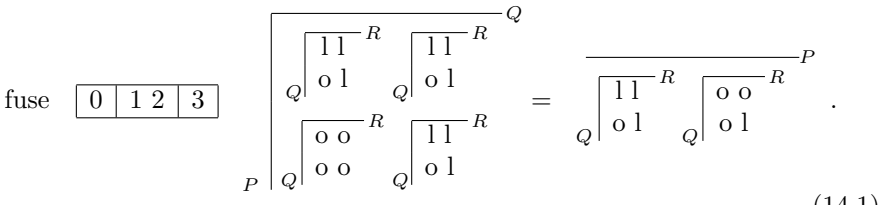
Picking a diagonal is a simple operation. Consider a function of n variables. If we consistently use the same variable for two (or $m+1$) different arguments, the result is a function of $n-1$ (or $n-m$) arguments. This geometrical image of the resulting function is a (hyper)diagonal in the geometrical image of the original function. In a tensor or array context, picking a diagonal is merely the operation of setting indices equal. While the operation is often considered to be too simple to mention, it is very important in the array-theoretic approach. Charles Sanders Peirce recognised the importance of the operation and referred to it as *colligation* [Peirce 1960; Franksen and Falster 2000]. The operation for picking diagonals is called **fuse** in array theory [More 1981].

Mike Jenkins [1981] developed a system called *Nial* (*N*ested *i*nter*a*ctive *l*anguage) for testing array-theoretic concepts. Let us follow the definition of **fuse** used in Nial. The operation **fuse** is used for two distinct purposes [Nial 2006]: (1) To perform a permutation of the axes in an array and (2) to pick diagonals between axes. Let I denote an array holding an index for each axis in the array A, then

$$\text{fuse } I \ A$$

returns the diagonal between the axes with indices that are grouped together (nested) in I. The axes are ordered according to the ordering in I. Take the four-dimensional array (13.5) from Section 13.1 which involves the variables P, Q, and R as an example. We fuse the two axes which both represent Q as

follows:

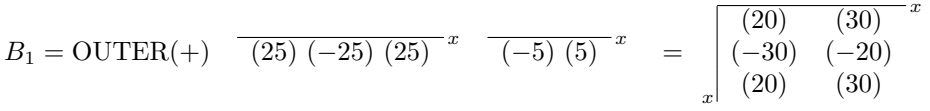


(14.1)

This process is exactly the same for expanded continuous arrays. However, when we consider an isomorphic representation, the entire diagonal may no longer be stored explicitly in the arrays. This means that we may need an operation which is different from **fuse** to find the diagonal in an isomorphic representation.

In his compact isomorphic representation, Møller [1995] uses arguments for cartesian products to point out the values of truth in a Boolean-valued array. The operation that performs colligation of these cartesian arguments is based on the set operation *intersection*. The polynomial array was introduced as an isomorphic representation of continuous arrays in Chapter 13. Polynomial arrays also require an operation for colligation which is different from **fuse**. The operation for colligation of polynomial arrays is different depending on whether we want to colligate after an outer product or after an outer sum or difference. The outer sums and differences are simpler because they do not change the degree of the Bézier surface. The operation **polyfuseP**, which was introduced in Section 13.3, is for colligation of redundant axes in outer *product* arrays. In the following, colligation of outer sums and differences is addressed first, and the operation **polyfuseP** is defined subsequently.

To explain the operation needed for colligation of axes after an outer sum or an outer difference, we use an example. Let us construct the polynomial array for $x^2 + x$ with $x \in [-5, 5]$. Using the array for x^2, which we found previously, we have



We would like to colligate the two axes in B which represent the same variable. Since outer sums and differences do not increase the degree of a (hyper)surface, one should expect that the traditional **fuse** operation would be appropriate for colligating repeated axes. Indeed the traditional **fuse** is the right operation. However, the two axes that we need to colligate must be of equal length. This problem is not difficult to solve, because we can raise the degree of an arbitrary Bézier curve without changing the curve.

To raise the degree of an arbitrary Bézier curve, we use the *degree elevation*

theorem. If $\boldsymbol{b}_0, \ldots, \boldsymbol{b}_n$ is a list of control points, the degree elevation theorem says that the following control points describe the same curve, but using one extra control point [Gravesen 2002]:

$$\begin{aligned} \hat{\boldsymbol{b}}_0 &= \boldsymbol{b}_0 \\ \hat{\boldsymbol{b}}_k &= \frac{n+1-k}{n+1}\boldsymbol{b}_k + \frac{k}{n+1}\boldsymbol{b}_{k-1} \quad , \quad k = 1, \ldots, n \\ \hat{\boldsymbol{b}}_{n+1} &= \boldsymbol{b}_n \ . \end{aligned}$$

This is not difficult to implement in array theory. Let us call the operation **raisedegree**. It is defined in Appendix B.1. This operation works with a list of control points, but we would like to elevate one dimension of an n-dimensional tensor product Bézier surface. Let us use the example to illustrate how this is done.

In the example the last axis is too short. If we nest this axis using **polynest**, we get the following Bézier curve:

$$\overline{\boxed{(20)\ (-30)\ (20)} \mid \boxed{(30)\ (-20)\ (30)}}^{\,x} \ .$$

If we elevate the degree of this curve, we have all the control points we need for the original array B. The result is:

$$\overline{\boxed{(20)\ (-30)\ (20)} \mid \boxed{(25)\ (-25)\ (25)} \mid \boxed{(30)\ (-20)\ (30)}}^{\,x} \ .$$

To remove the nesting and return to the two-dimensional array, there is an operation in array theory called **blend** I A which removes one level of nesting and places the axes of the items in A at the top level in the order given by I. Since **polynest** operates on the last axis of the array, we give all the first axes as the argument (I) to blend. The general function which elevates the degree of the last dimension is defined by

$$\text{elevate} = \text{blend}\ [\text{front axes}, \text{raisedegree polynest}] \ .$$

In the example we get

$$B_2 = \text{elevate}\ B_1 \ = \ {}_{x}\left|\begin{array}{ccc} (20) & (25) & (30) \\ (-30) & (-25) & (-20) \\ (20) & (25) & (30) \end{array}\right|^{x} \ .$$

After elevation such that both x-axes have the same length, the colligated array is obtained using the conventional **fuse**:

$$B_{x^2+x} = \text{fuse}\ (0\ 1)\ B_2 = \ \overline{(20)\ (-25)\ (30)}^{\,x} \ .$$

Again it is not difficult to verify that this is the correct polynomial array for $x^2 + x$. Suppose we had wanted to find the array for $x + x^2$, then the axis to be elevated had not been the last axis of the array. The operation **elevate** was, however, defined always to work on the last axis. This is not a problem, since we can always rearrange the axes in an array using the operation **fuse**.

Now, let us define the **polyfuseP** operation which is needed for colligating repeated axes in polynomial arrays that result from an outer product. A method that determines the control points for the diagonal of a tensor product Bézier surface has been described by Holliday and Farin [1999]. The traditional tensor product Bézier surface corresponds to the two-dimensional case in terms of a polynomial array. In the following we will generalise the approach of Holliday and Farin such that we are able to find the diagonal between an arbitrary number of axes in a multidimensional array. The structural operations of array theory helps tremendously in this generalisation.

To facilitate the definition of the operation **polyfuseP**, we will use an example. Let us use the polynomial array for

$$(x - x_0)^2 \quad , \quad x, x_0 \in [-5, 5] \ ,$$

as an example. Before colligation the corresponding polynomial array is

$$A \quad = \quad \begin{array}{|c|c|}\hline \begin{array}{|c|c|}\hline (0) & (0) \\ (0) & (0) \\ \hline \end{array}_{x} & \begin{array}{|c|c|}\hline (0) & (100) \\ (-100) & (0) \\ \hline \end{array}_{x} \\ \hline \begin{array}{|c|c|}\hline (0) & (-100) \\ (100) & (0) \\ \hline \end{array}_{x} & \begin{array}{|c|c|}\hline (0) & (0) \\ (0) & (0) \\ \hline \end{array}_{x} \\ \hline \end{array}_{x_0} \quad .$$

This array is obtained by an outer difference of the scales of x and x_0 and an outer product of the resulting array with itself. The proposed operation works as follows. Let I denote an array holding the indices of the axes in the array A that are to be colligated. The first step is to take the **grid** of A.

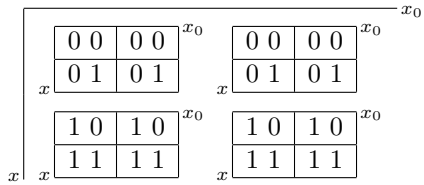
$$\text{grid } A \quad = \quad \begin{array}{|c|c|}\hline \begin{array}{|c|c|}\hline 0\,0\,0\,0 & 0\,0\,0\,1 \\ 0\,0\,1\,0 & 0\,0\,1\,1 \\ \hline \end{array}_{x} & \begin{array}{|c|c|}\hline 0\,1\,0\,0 & 0\,1\,0\,1 \\ 0\,1\,1\,0 & 0\,1\,1\,1 \\ \hline \end{array}_{x} \\ \hline \begin{array}{|c|c|}\hline 1\,0\,0\,0 & 1\,0\,0\,1 \\ 1\,0\,1\,0 & 1\,0\,1\,1 \\ \hline \end{array}_{x} & \begin{array}{|c|c|}\hline 1\,1\,0\,0 & 1\,1\,0\,1 \\ 1\,1\,1\,0 & 1\,1\,1\,1 \\ \hline \end{array}_{x} \\ \hline \end{array}_{x_0} \quad .$$

In our example we have two colligations to do, because there are two pairs of axes representing the same variable. Let us begin by colligating the x-axes, then $I = (0\ 2)$. From each item in **grid** A, we choose the coordinates given by I. Let us call this operation **polychoose**. Array-theoretically it is defined by

$$\text{polychoose } I \ A = \text{EACH}(I \text{ choose}) \text{ grid } A \ ,$$

where the operation I **choose** A picks the items in A at the positions given by the indices in I. The following is the result in our example:

polychoose I A =



.

The next step is to sum the items of each of these nested arrays and link the summed coordinate with the remaining coordinates in the grid. Let us call this operation **polyindex**. To link the summed coordinate with the remaining coordinates in the grid, we need an operation to find the indices of the axes that are not in I. For this purpose we introduce the operation

$$\text{otheraxes} = \text{except}\ [\text{axes second}, \text{first}]\ ,$$

where **axes** , as mentioned before, lists the indices of the axes in the array given as argument, and **first** and **second** pick the first argument and the second argument given to **otheraxes**. The operation **except** J I removes the items from J which are also in I. In our example, we have

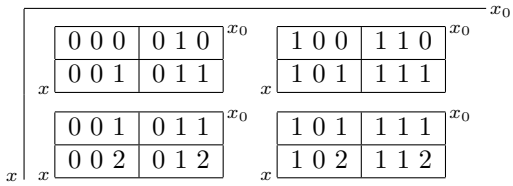
$$\text{otheraxes}\ I\ A = (1\ 3)\ .$$

With the operations **polychoose** and **otheraxes**, we define **polyindex** by

$$\begin{aligned}\text{polyindex} \quad &= \quad \text{EACH(link) pack} \\ &\qquad [\text{polychoose}\ [\text{otheraxes}, \text{second}], \text{EACH}(+)\ \text{polychoose}]\ ,\end{aligned}$$

where **link** turns all the items given as argument into a list of items. Again looking at the example, the result of **polyindex** is

polyindex I A =

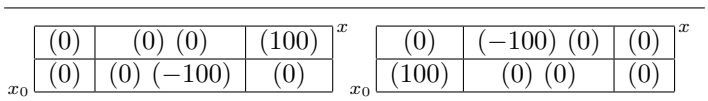


.

The coordinate indices in this array correspond to the positions in a new three-dimensional array. Let us introduce a new operation called **polyplace** which places the items of A at the positions given by **polyindex** I A in a new array. The result is

polyplace I A

=



.

The array-theoretic definition of **polyplace** uses **polyindex**, but it is a little lengthy, so the definition is given in Appendix B.2. Intuitively it is not difficult to figure out how the operation works by inspecting the example.

It should be observed that several control points are nested at the same location in the new array obtained using **polyplace**. In Section 13.3, where the operation **polyfuseP** was proposed for picking the diagonal of polynomial arrays, we did not discuss how to handle these nested control points. The reason why it was not discussed is that the nested control points were the same in the examples we considered. This is, however, not true in general. The general way of handling the nested control points is not obvious. For a moment, let us call the operation to handle them P_f. The definition of **polyfuseP** is then

$$\text{polyfuseP} = \text{EACH}(P_f)\ \text{polyplace}\ .$$

Using the technique that finds the control points for the diagonal curve of a tensor product Bézier surface [Holliday and Farin 1999], we are able to define the operation P_f. The technique is actually quite simple. Each nested Bézier curve is degree elevated until its degree is doubled. This means that if it has n control points, it is degree elevated $n-1$ times. The middle control point in the elevated list of control points is the control point that belongs to the diagonal. Let us define P_f using array theoretic operations.

First we need an operation to pick the middle control point in a list of control points. This is defined as follows:

$$\text{middle} = \text{pick}\ [\text{floor}\ (2\ \text{CONVERSE}(/)\ \text{shape}), \text{pass}]\ ,$$

where the operation / denotes division and the operation **floor** returns the integral part of a floating point number. With the operation **middle**, we are ready to define P_f. This is done using the FOLD operator described previously (13.7). We have

$$P_f = \text{middle FOLD}(\text{raisedegree})\ [-1 + \text{tally}, \text{pass}]\ ,$$

where **tally** A returns the number of items in the array A. This concludes the definition of the operation **polyfuseP**.

In our example, the result is

$$\text{polyfuseP}\ I\ A = \begin{array}{c} \overline{}^{x_0} \\ \left\lceil \begin{matrix} (0) & (0) & (100) \\ (0) & (-50) & (0) \end{matrix} \right\rceil_{x_0}^{x} \quad \left\lceil \begin{matrix} (0) & (-50) & (0) \\ (100) & (0) & (0) \end{matrix} \right\rceil_{x_0}^{x} \end{array} .$$

When we colligate the two x_0-axes, again using the operation **polyfuseP**, we

obtain:

$$\text{polyfuseP } (0\ 1)\ (\text{polyfuseP } (0\ 2)\ A) = \begin{array}{|ccc|}\hline (0) & (0) & (100) \\ (0) & (-50) & (0) \\ (100) & (0) & (0) \\ \hline \end{array}\ {\scriptstyle x_0 \atop x} \quad .$$

If we had begun with $x, x_0 \in [a, b]$ as in Chapter 13, we would have arrived at the same formula as the one we got there (13.10). Now the general idea has been described more precisely.

Conclusively, polynomial arrays need two different operations for colligation. One for colligation after an outer product of two polynomial arrays (**polyfuseP**) and one for colligation after an outer sum or difference of two polynomial arrays (**fuse**, and **elevate** if the length of the axes to be colligated do not match). These are operations which we can use to remove redundant axes which represent the same variable after an array has been created using outer transforms.

14.3 Extraction

So far this part of the thesis has mostly been concerned with representation of multidimensional shapes. When a geometrical representation has been obtained, we are ready to learn about the shape. In other words, we are ready to use the appearance model.

An ordinary way in which to use an appearance model is to choose a specific value for some input, and then to see how the output behaves when the remaining input variables are varied. We can easily accomplish this when we have an array representation using nesting and picking. For a regular grid representation of continuous arrays, we would use the operations **split** and **pick**. For polynomial arrays, we would use **polynest** and **polypick**. The additional advantage of the geometrical approach is that we are also able to choose specific values for some output, and then see how the input behaves. The distinction between input and output is not explicit in the arrays. Both an input and an output is just another axis. What I mean by input and output for an appearance model was explained in Chapter 12.

Another opportunity, which the geometrical approach provides us with, is to find the general relation between a subset of the involved variables. A relation which is true regardless of the value of the remaining variables. In other words, the geometrical approach makes us able to derive the relations between variables that the shape implies. Such relations are found by *orthogonal projection* of the geometrical image on a slice of the multidimensional space which is spanned by

a subset of the involved variables. Thus orthogonal projection is as fundamental in the handling of shapes as it is in the handling of systems of quantum particles.

When we have an array representation, orthogonal projection is accomplished using nesting. To start with a simple example, consider (again) the constraints:

$$\begin{aligned} C_1 &: P \Rightarrow Q \\ C_2 &: Q \Rightarrow R \ , \end{aligned}$$

where P, Q, and R are Boolean variables. These constraints describe a shape. Using OUTER and **fuse**, we find that the array C, which represents the shape, is

$$C = \overline{\begin{array}{cc} \overline{\begin{array}{|c|}\hline 1\ 1 \\ \hline o\ 1 \\ \hline \end{array}}^{R} & \overline{\begin{array}{|c|}\hline o\ o \\ \hline o\ 1 \\ \hline \end{array}}^{R} \\ {\scriptstyle Q} & {\scriptstyle Q} \end{array}}^{P} \ .$$

Suppose we want to find the relation between P and R disregarding Q. Then all we have to do is an orthogonal projection of the Q-axis in the array.

To nest the Q-axis, we use the operation **split** which was introduced in Section 13.3. The result is

$$\text{split } 1\ C = \overline{\begin{array}{|c|c|}\hline 1\ o & 1\ 1 \\ \hline o\ o & o\ 1 \\ \hline \end{array}}^{R}_{P} \ .$$

Intuitively, this split means that every nested array holds what we are seing in the direction along the Q-axis. Thus if there is a single value of truth in a nested array, there should be a value of truth at the position where the nested array is. This corresponds to logical disjunction between all the items of each nested array. Let us call the operation **project**, it is defined by

$$\text{project} = \text{EACH}\,(\text{REDUCE}(\vee))\ \text{split} \ .$$

In the example, the result is

$$\text{project } 1\ C = \overline{\begin{array}{|c|}\hline 1\ 1 \\ \hline o\ 1 \\ \hline \end{array}}^{R}_{P} \ .$$

Normally the projected geometrical image is all we need. We can use it as a look-up table which describes the relation between P and R. The projected relation is valid no matter what the value of Q is. It is not so difficult to prove that the relation between a number of variables, which is described by a shape, implies all the relations given by orthogonal (disjunctive) projections on an arbitrary space spanned by a subset of the involved variables. This means that orthogonal projection provides a mechanical way of proving theorems. In fact, we have just found a mechanical proof of Aristotle's famous hypothetical

syllogism. The relation we found between P and R is evidently $P \Rightarrow R$, thus we have

$$\frac{\begin{array}{c} P \Rightarrow Q \\ Q \Rightarrow R \end{array}}{P \Rightarrow R} \ ,$$

where the horizontal line means that an implication between the constraints above the line and the conclusion below the line is a tautology. A branch of logic called *array-based logic* was founded by Franksen [1979] on the three fundamental principles that have been presented in this chapter. The three principles are outer transforms for creation or expansion of the geometrical representation (Sec. 14.1), colligation for reduction of the geometrical representation (Sec. 14.2), and orthogonal projection as well as picking of subspaces for extraction of relations from the geometrical representation (Sec. 14.3). These principles provide a simple way of doing mechanical reasoning. While reasoning is not a key subject in this thesis, it is interesting to note the broad applicability of these three fundamental principles. From reasoning to quantum mechanics to appearance modelling.

In a regular grid representation of a continuous array, the operation for projection is the same as for the Boolean arrays. Of course, the result will only be approximate since the regular grid corresponds to a sampling of the continuous array. A polynomial array is, on the other hand, a fit of the continuous array using a Bézier surface with one-dimensional control points. If the constraints are simple mathematical formulae, we are able to obtain a precise fit. This is obtained by outer transforms as discussed in Sections 13.3 and 14.1. More commonly each constraint, which is part of an appearance model, will be measured data or the result of complicated calculations such as evaluation of the Lorenz-Mie formulae. In this case, a fit of control points to each constraint is needed. These fits can be obtained using a standard mathematics tool such as Maple or, for example, as described by Cohen et al. [2001]. When the fits for the different constraints have been found, they are combined into a single polynomial array using the rules described in Section 14.1. All redundant axes resulting from outer sums, and all redundant axes resulting from multiplication of linear fits, should be eliminated using the colligation rules discussed in Section 14.2. In this way, we obtain a single polynomial array which is the geometrical representation of our appearance model.

As mentioned previously, the polynomial array is an implicit surface. If we sample the values in the array using the operation **polypick**, we can transform it into a regular grid representation of arbitrarily fine resolution. The surface is given by the convention which is usually EACH(0 =). Compared to use of sampled continuous arrays throughout the creation of the geometrical representation, the polynomial array is a much more compact and precise solution. When the polynomial array has been converted into a sampled rectangular grid

representation of arbitrary precision, we can project it orthogonally using the operation **project**. Or we can use any other graphics technique for orthogonal projection of an implicit surface. Indeed orthogonal and perspective projections are one of the most well-developed techniques in graphics. Such projections are the very foundation of rendering, where we project a scene on an image plane.

14.4 Conclusions

The general purpose of this part has been to provide a mathematical framework for building versatile appearance models. In particular, it has been argued that the geometrical representation of appearance models gives several advantages. The strongest advantage is that we can freely vary all input and output. The input we are referring to are the contents of a material, the particle sizes, the indices of refraction, and any physical conditions they may depend on. The output are the macroscopic optical properties. The free variation property of the geometrical representation enables us not only to model how the appearance of a material changes as we change the input parameters, but also to model how the input parameters change as we change the macroscopic optical properties.

In the conclusion of Part II (Sec. 11.2), we compared the suggested theory for computation of macroscopic optical properties to measurement of optical properties. The conclusion was that the two methods are complementary rather than competitors. The idea of a geometrical representation of appearance models, clearly demonstrates that the two techniques are complementary. Because with measured macroscopic optical properties, we have a way of coupling appearance to the output of our appearance model, and the geometrical representation enables us to analyse the influence of the macroscopic optical properties on the input for the appearance model. Thus we also have a way of *analysing* appearance. Through this type of analysis, we can learn about the meaning of the appearance of materials as discussed in Chapter 12. This completes our theoretical discussion of light, matter, and geometry. In the next part we will construct three appearance models.

Part IV

APPEARANCE

Chapter 15

Water

It was the only color we could see in the universe.

Bill Anders, Apollo 8 astronaut

The theories in the first three parts of this thesis describe how to bring out the relation between the microscopic properties of a material and the appearance of the material. All we need, to get started, are some measurements describing the microscopic properties. The purpose of this fourth part is to exercise the theory. To do this, we investigate measurements of microscopic properties (Sections 15.1, 16.1, 17.1). Then we create appearance models, which link the microscopic properties to the macroscopic properties as well as the physical conditions of the materials (Sections 15.2, 16.2, 17.2). Finally, we use the appearance models to render the considered materials under different circumstances (Sections 15.3, 16.3, 17.3). This makes us able to draw conclusions about the influence of each ingredient - each different particle inclusion - on the appearance of the materials.

Before we construct the appearance models, let us first outline the general procedure. There are three steps (1.-3.). The presented theory requires some input. To compute the macroscopic optical properties for a material, we need shapes, complex indices of refraction, and size distributions for the different particle types in the material. We also need the complex index of refraction

of the host medium. Therefore the first step (1.) is to seek information about the material. Indices of refraction (including the imaginary part) are often available in chemical software packages (e.g. PhotochemCAD [Du et al. 1998]), physics handbooks [Gray 1972; Palik 1985; Lide 2006], and online repositories (e.g. http://www.luxpop.com/). Particle shape and size distributions can often be found in optics or chemistry literature, or some mean particle size can be estimated. Once the input parameters have been obtained for one substance, they may be used in many different contexts (in the following we will use the properties of water several times and the properties of minerals and algae for both water and ice).

Unfortunately the input concerning specific particle types is not a very intuitive set of parameters. Therefore it is desirable to find a set of input from which the particle specific input can be derived. This type of input is typically volume fractions for the different particle types and physical conditions such as temperature. The second step (2.) is to find a desirable set of input parameters. This step is entirely dependent on the amount of information that we can find about the material. Usually indices of refraction have some dependency on temperature, and sometimes volume fractions are convenient.

The third step (3.) is to compute the optical properties. The procedure for computing the optical properties is provided in Section 15.2. A recommended option is to compute sample optical properties for different input, and to fit a surface to the result. Then we can use the fit as a constraint for a geometrical representation of the appearance model. This concludes the outline of the procedure. Now, let us construct an appearance model for water.

Deep waters would appear black (apart from surface reflection) if they did not contain scattering particles. A real-world example of the colour caused by mineral sediments in water is Lake Pukaki, New Zealand, where glacial melt water with a high concentration of "rock flour" mixes with clear water from melted snow. The result is an impressive blue colour, see Figure 15.1.

Natural waters contain many different particle inclusions, but to keep this first example simple, we model only the two most visually significant types of particles. Algae are visually significant, so we include algae as one type of particle in our model. Minerals are also visually significant (as Figure 15.1 shows), but there are many different types of minerals. Following Babin et al. [2003a], we treat all types of minerals as being one and the same type of particle. In the following section we collect the optical properties of the host and of these two types of particles.

Figure 15.1: *Photograph of Lake Pukaki, New Zealand, which illustrates the influence of mineral suspensions on the colour of water.*

$n_1 = 1.779 \cdot 10^{-4}$	$n_4 = -2.02 \cdot 10^{-6}$	$n_7 = -0.00423$
$n_2 = -1.05 \cdot 10^{-6}$	$n_5 = 15.868$	$n_8 = -4382$
$n_3 = 1.6 \cdot 10^{-8}$	$n_6 = 0.01155$	$n_9 = 1.1455 \cdot 10^{6}$

Table 15.1: *The coefficients for the empirical formula* (15.1) *by Quan and Fry* [1995].

15.1 Particle Composition

The host medium is pure water or brine (saline water). It is well-known that water is transparent in small quantities and blue in large quantities. This means that pure water is a weakly absorbing host medium, and the imaginary part of the index of refraction is not neglegible.

Quan and Fry [1995] have found an empirical formula for computing the real part of the refractive index of pure water or brine as a function of salinity S, temperature T, and wavelength λ. It is as follows [Quan and Fry 1995]:

$$\begin{aligned} n'_{\text{water}}(\lambda, T, S) &= 1.31405 + (n_1 + n_2 T + n_3 T^2) S + n_4 T^2 \\ &\quad + \frac{n_5 + n_6 S + n_7 T}{\lambda} + \frac{n_8}{\lambda^2} + \frac{n_9}{\lambda^3} \ . \end{aligned} \tag{15.1}$$

The coefficients are listed in Table 15.1. This formula describes the dependency

λ [nm]	n''_{water}	Ψ_T	Ψ_S
375	$3.393 \cdot 10^{-10}$	0.0001	0.00012
400	$2.110 \cdot 10^{-10}$	0.0001	0.00012
25	1.617	0.00005	0.000055
50	3.302	0.0	−0.00002
75	4.309	0.0	−0.00002
500	$8.117 \cdot 10^{-10}$	0.0001	−0.00002
25	$1.742 \cdot 10^{-9}$	0.0002	−0.000025
50	2.473	0.0001	−0.00003
75	3.532	0.0002	−0.00002
600	$1.062 \cdot 10^{-8}$	0.0010	−0.000015
25	1.410	0.0005	−0.00001
50	1.759	0.00005	0.0
75	2.406	0.0001	−0.00002
700	$3.476 \cdot 10^{-8}$	0.0002	−0.00017
25	8.591	0.0065	−0.00001
50	$1.474 \cdot 10^{-7}$	0.0106	0.00064
775	$1.486 \cdot 10^{-7}$	0.0106	0.00064

Table 15.2: *Imaginary part of the refractive index for pure water* n''_{water} *collected from Pope and Fry* [1997] ($375\,\text{nm} \leq \lambda \leq 700\,\text{nm}$) *and Hale and Query* [1973] *(remaining wavelengths). Correction coefficients,* Ψ_T *and* Ψ_S*, from Pegau et al.* [1997] *are also included to make the indices useful for both brine and water at various temperatures. Values in this table which do not appear in the references are interpolations.*

of n'_{water} on salinity in the range $0‰ < S < 35‰$, temperature in the range $0\,°\text{C} < T < 30\,°\text{C}$, and wavelength in the range $400\,\text{nm} < \lambda < 700\,\text{nm}$. Moreover Huibers [1997] has reported that the same formula is valid over a broader spectrum of wavelengths ($200\,\text{nm} < \lambda < 1100\,\text{nm}$) than originally assumed.

The imaginary part of the refractive index has most recently been measured by Pope and Fry [1997]. They measured it for the wavelengths $380\,\text{nm} < \lambda < 700\,\text{nm}$. Measurements for the remainder of the visible spectrum are available from Hale and Query [1973]. These measurements of the imaginary part of the refractive index for pure water are collected in Table 15.2. The dependency of the imaginary part on temperature and salinity has been measured in terms of a set of correction coefficients by Pegau et al. [1997]. The correction formula is

$$n''_{\text{water}}(\lambda, T, S) = n''_{\text{water}}(\lambda, T_r, 0) + \lambda \frac{(T - T_r)\Psi_T + S\Psi_S}{4\pi} \quad ,$$

λ [nm]	n''_{mineral}	n''_{alga}
375	$2.12 \cdot 10^{-3}$	$1.84 \cdot 10^{-4}$
400	$1.66 \cdot 10^{-3}$	$1.96 \cdot 10^{-4}$
25	1.30	2.89
50	1.01	3.11
75	$7.84 \cdot 10^{-4}$	2.79
500	$6.07 \cdot 10^{-4}$	$2.14 \cdot 10^{-4}$
25	4.59	1.26
50	3.61	$8.19 \cdot 10^{-5}$
75	2.77	5.57
600	$2.13 \cdot 10^{-4}$	$6.03 \cdot 10^{-5}$
25	1.63	7.86
50	1.25	$1.00 \cdot 10^{-4}$
75	$9.52 \cdot 10^{-5}$	2.53
700	$7.26 \cdot 10^{-5}$	$3.91 \cdot 10^{-5}$
25	5.53	3.91
50	4.20	3.91
775	$3.19 \cdot 10^{-5}$	$3.91 \cdot 10^{-5}$

Table 15.3: *Imaginary part of the refractive index for minerals and algae. These spectra are typical examples computed using the empirical formulae by Babin et al. [2003b] (minerals) and Bricaud et al. [1995] (algae). The empirical formulae depend on the density of the minerals and the concentration of algae respectively.*

where Ψ_T and Ψ_S are the correction coefficients listed in Table 15.2 and T_r is the reference temperature. The measurements of Pope and Fry [1997] were carried out at the temperature $T_r = 22℃$.

This was quite a lot of effort to describe the refractive index of water as a function of temperature and salinity. Water and brine are, however, useful in many contexts, so the effort has not been wasted on a single example.

Now that we know the optical properties of the host medium, let us look at the particle inclusions. Since the different types of minerals are modelled as one type of particle, we use an approximate index of refraction, the real part is [Babin et al. 2003a]

$$n'_{\text{mineral}} = 1.58 \ .$$

An empirical formula for computing the imaginary part of the refractive index for minerals has been found by Babin et al. [2003b]. A typical spectrum is listed

Property	Pure water (host)	Mineral	Alga
n'	Equation 15.1	1.58	1.41
n''	Table 15.2	Table 15.3	Table 15.3
r		$[0.01\,\mu\text{m}, 100\,\mu\text{m}]$	$[0.225\,\mu\text{m}, 100\,\mu\text{m}]$
$N(r)$		Equation 15.2	Equation 15.3

Table 15.4: *A summary of the microscopic properties of natural waters.*

in Table 15.3. Algae approximately have the following real part of the refractive index [Babin et al. 2003a]:

$$n'_{\text{alga}} = 1.41$$

and an empirical formula for the imaginary part of the refractive index has been found by Bricaud et al. [1995]. A typical spectrum for algae has also been listed in Table 15.3.

The mineral and algal particles are assumed spherical. Number densities follow the power law (10.3). Particle radii should be integrated over the intervals $r_{\text{mineral}} \in [0.01\,\mu\text{m}, 100\,\mu\text{m}]$ and $r_{\text{alga}} \in [0.225\,\mu\text{m}, 100\,\mu\text{m}]$. The power laws have been estimated for particle diameters measured in μm. Using the relation between number density and volume fraction (10.4), we find

$$N_{\text{mineral}}(r) = \frac{v_{\text{mineral}}}{10.440}(2r)^{-3.4} \tag{15.2}$$

$$N_{\text{alga}}(r) = \frac{v_{\text{alga}}}{4.9735}(2r)^{-3.6}\,, \tag{15.3}$$

where v_{mineral} and v_{alga} are the volume fractions of minerals and algae in the water, respectively, and radii r are measured in μm such that the result is number density measured in μm^{-4}. The powers $\alpha = 3.6$ and $\alpha = 3.4$ have been reported by Babin et al. [2003a]. The microscopic properties of natural water are summarised in Table 15.4.

15.2 Appearance Model

Assuming that the microscopic properties of minerals and algae do not change significantly with the temperature or the salinity of the water they are in, the information provided in the previous section maps the following parameters, which concern a sample of water, to the microscopic properties of the water:

- Temperature T
- Salinity S
- Volume fraction of minerals v_{mineral}
- Volume fraction of algae v_{alga} .

Going from microscopic to macroscopic properties is accomplished using the theory described in Part II. The steps are as follows:

1. Number density of mineral and algal particles are determined such that they constitute the desired volume fractions (Equations 15.2 and 15.3, which are based on Equations 10.4 and 10.3).

2. For every wavelength sampled, and for every particle type and particle radius considered, we evaluate the Lorenz-Mie coefficients a_n and b_n (Equations 9.18, 9.14, 9.19, 9.15, 9.16, 9.17, 9.12, and 9.13), and we compute the extinction cross section C_t, scattering cross section C_s, and asymmetry parameter g_p (Equations 9.21, 9.22, 9.23, and 9.24).

3. Scattering coefficient $\sigma_{s,i}$, extinction coefficient $\sigma_{t,i}$, and combined asymmetry parameter g_i are determined for each particle inclusion $i \in \{\text{mineral, alga}\}$ (Equations 10.1 and 10.9).

4. Finally bulk extinction coefficient σ_t, bulk scattering coefficients σ_s, bulk absorption coefficient σ_a, ensemble asymmetry parameter g, and effective refractive index are computed (Equations 10.5, 10.6, 10.11,and 10.7).

This procedure works for arbitrary volume fractions of mineral and algal particles, and it is the same for other materials if we swap the mineral and algal particles for a different set of particles. Thus we have a four-dimensional appearance model for natural waters.

The temperature and salinity of ocean waters throughout the world is mapped continually in the World Ocean Atlas Series and the data is freely available [Locarnini et al. 2006; Antonov et al. 2006]. The properties of pure water described in the previous section cover almost the entire range of temperatures and salinities found throughout the world (except for freezing brine, which is described in Section 16.2). This means that our appearance model captures most natural waters on the face of the Earth.

Region	v_{alga}	v_{mineral}
Atlantic	$1.880 \cdot 10^{-7}$	$2.477 \cdot 10^{-10}$
Baltic	$1.904 \cdot 10^{-6}$	$5.429 \cdot 10^{-7}$
Channel	$4.999 \cdot 10^{-7}$	$2.300 \cdot 10^{-7}$
Mediterranean	$3.878 \cdot 10^{-7}$	$3.075 \cdot 10^{-7}$
North Sea	$2.171 \cdot 10^{-6}$	$2.077 \cdot 10^{-6}$

Table 15.5: *Volume fractions of algae and minerals contained in coastal and oceanic waters around Europe. Data reported by Babin et al.* [2003a] *have been translated into the volume fractions given here.*

Potential Expansion of the Model

Suppose we model the ocean as a plane-parallel water slab of thickness d on an opaque sand coloured background. Then, given a set of bulk optical properties for the water, it is not difficult to render the colour of the ocean as a function of d. Doing such simple renderings, we are able to obtain a six-dimensional shape describing the relation between T, S, v_{mineral}, v_{alga}, d, and colour.

Another shape we can make is one which describes the constraints given by a water depth map of the globe (note that the depth corresponds to d), a water temperature map of the globe, and a water salinity map of the globe. This is a five-dimensional shape describing the relation between d, T, S, and u, v positions on the globe. An OUTER($\wedge$) of these two shapes, a colligation of the axes representing d, T, and S, and an orthogonal projection of the axes representing d, T, and S results in an appearance model describing the relation between v_{mineral}, v_{alga}, u, v positions on the globe, and the colour of the sea. This is an incredibly powerful appearance model. Suppose we are on a boat with a camera. We take a picture of the water (perhaps close to the surface to eliminate surface reflection), give it as input to a program along with our current position on the globe. Then, using the appearance model, the program is able to compute the range of possible mineral and algal contents of the water. Or perhaps we do not know our position, but we are able to measure the contents of the water. Then the program is able to give us the set of possible positions on the globe that we might be at.

Other interesting appearance models could be constructed, if we had global measurements of mineral and algal contents of ocean water. I have not been able to find this. Babin et al. [2003a] have sampled the contents of coastal and oceanic waters around Europe. Volume fractions based on these measurements are given in Table 15.5.

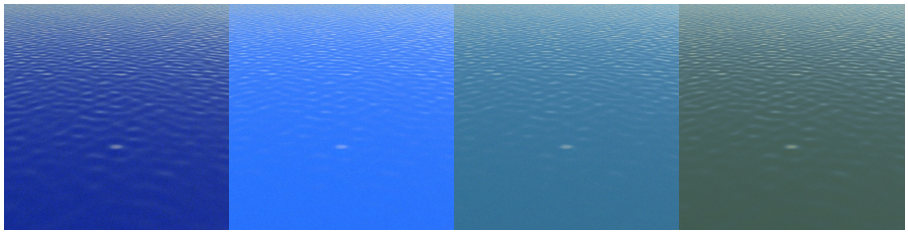


Figure 15.2: *Coastal and oceanic waters with the mineral and algal contents given in Table 15.5.*

15.3 Results

Figure 15.2 shows rendered images of four different oceanic and costal waters (modelled as being very deep). The waters are all illuminated by the same atmospheric lighting model and the color difference in the images are only due to the optical properties of the water. The water ripples have been generated procedurally. Note that a larger amount of minerals gives the water a lighter blue colour. The blue colour is due to the absorption of the pure water host. As the amount of scattering mineral particles increases, the amount of light being scattered back from the water increases. A larger amount of algae makes the water more green. This is easily observed in the rendering of the North Sea. The greener colour is due to the absorption spectrum (the imaginary part of the refractive index) of the algal particles.

In Figure 15.3 the rendered appearance of the different oceanic and costal waters have been placed in a map drawn by Babin et al. [2003a] to show where they took the samples of the mineral and algal contents of the water. This concludes the water example. In the next chapter, we will look at ice as an example. While ice and water are closely related chemically, they are quite different in the forms found in nature. Note that ice is a solid. Thus the ice example is a way of demonstrating that the theory also works for solids.

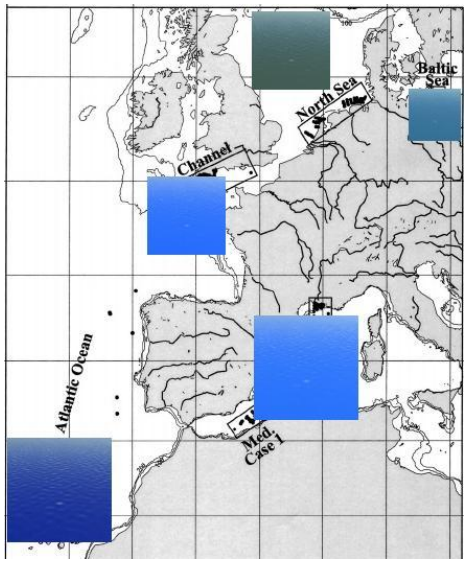


Figure 15.3: *The rendered images from Figure 15.2 have been placed in the map showing the locations where the mineral and algal contents of the water was measured by Babin et al.* [2003a].

Chapter 16

Ice

There is also ice of a different shape which the Greenlanders call icebergs. In appearance these resemble high mountains rising out of the sea; they never mingle with other ice but stand by themselves.

Unknown author of *The King's Mirror* (c. 1250)
– a textbook written for the sons of King Magnus Lagabøte

Ice in itself is a pure substance, but in reality we do not see it in its pure form. Even if we make ice cubes in a freezer using clean tap water, the result is not pure ice, it is ice "contaminated" by scattering air particles. This is the reason why ice cubes are rarely transparent. Why, then, do icebergs appear in many different colours in nature? Their colours range from intense white over deep blue and green to dark grey. These differences in colour appear even if the icebergs are observed under the same lighting conditions. The reason for this large variation in iceberg appearance is hidden in the composition of the embedded particles and in the optical properties of the pure ice host.

The reasons for the various appearances of ice have previously been discussed qualitatively (for example by Bohren [1983]). This provides an opportunity to test if our quantitative appearance models match the qualitative expectations. To facilitate such a comparison, the following is a recap of qualitative explanations for the appearance of ice.

Snow is perhaps the whitest substance found in nature and yet it can exhibit an incredibly pure blue color [Bohren 1983]. Considering the white surface of a snow covered landscape and taking into account that snow is composed of ice

grains, we are compelled to conclude that the surface of an ice medium reflects light non-selectively in the entire visible range of the spectrum. Nevertheless, if we dig a hole in a snowpack or look into a glacier crevasse, we are able to observe deep-blue light which exceeds the purity of the bluest sky [Bohren 1983]. The reason is that light is not only reflected off the surfaces of the ice grains, but also transmitted through them following the Fresnel equations (cf. Section 4.4). This tells us that pure ice, while not being spectrally selective with respect to light scattering, does indeed absorb light selectively in the visible range.

Because of the many ice grain surfaces on which light can reflect and refract, snow is highly scattering. Again, since snow is white and bright when we look at the surface, the scattering must be so frequent that light escapes without suffering from absorption. Of course, the probability of absorption when light is transmitted through an ice grain must also be small. Gradually, however, the further we penetrate into the snowpack, the more transmissions through ice grains will occur thus increasing the probability of absorption. Hence, the deeper we dig, the bluer the color. Since the transmitted color of snow is blue, we conclude that pure ice absorbs less light in the near-ultraviolet than in the near-infrared part of the visible range. All these observations about snow are, as we shall later discover, reflected in the optical properties of ice.

Pure ice is a hypothetical substance which does not occur in nature. Natural ice is cracked and bubbly. It is composed of air grains in ice as opposed to snow which is composed of ice grains in air. The bubbles occurring in ice are larger and less frequent than snow grains are in a snowpack. Hence, light travels through more ice and is less frequently scattered in ice as compared to snow. This means that the probability of absorption is larger when light travels in ice. Consequently, we do not have to dig a hole in an iceberg to see a blue color. If the iceberg has the right composition, such that scattering is infrequent, the iceberg is blue even if we only look at the surface.

Blue icebergs are most often observed in lakes or seas near glaciers. Glaciers compress the ice making the air bubbles collapse into the ice structure (the compound of ice and air is called a *clathrate* in chemical terms). When an iceberg comes loose from a well-compressed part of a glacier, it has few air bubbles and will appear blue or bluish. An iceberg grown in the sea (or sea ice, in short) has more scattering inclusions than fresh water ice. In particular, sea ice has vertically oriented brine pockets which come into existence as salt drains away from the freezing water. For this reason sea ice frequently has a whiter and more slushy appearance as compared to fresh water ice.

Soot, dust, soil, and organic particles also have an influence on the appearance of ice [Light et al. 1998]. This is the reason why icebergs with a greenish hue appear in nature. Another reason is that if a piece of ice is not very thick, the

underlying surface will have an influence on the appearance of the ice. Soot is black, soils and rocks are most often reddish or yellowish, most organic matter is green. Existence of any of these particulates in a piece of ice would shift the reflectance spectrum towards longer wavelengths [Bohren 1983]. The greenish appearance of some ice is, hence, either due to contamination of one sort or another or simply the result of some underlying surface.

In summary, we can observe white, bluish, and greenish icebergs under the same lighting conditions since all these different appearances are due to the composition of the scattering inclusions rather than the composition of the incident light. There is only one exception to this rule and that is the bottle green iceberg which we have not yet addressed. The bottle green iceberg is a rare phenomenon which seems to occur only under certain conditions [Kipfstuhl et al. 1992; Warren et al. 1993]. One theory is that such an iceberg is only bottle green under certain lighting conditions [Lee 1990]. Bottle green icebergs will be described as a special case later in this chapter.

16.1 Particle Composition

Pure ice is the host medium. Perhaps surprisingly it does not have the same index of refraction as pure water. They are similar, but not the same. Many researchers have endeavored to measure the real part and especially the imaginary part of the refractive index for pure ice. A compilation of these efforts up to 1984 has been given by Warren [1984]. In the visible range of the spectrum the real part of the refractive index is largely independent of temperature and wavelength, $n'_{\text{ice}} \approx 1.31$. It varies approximately 1% over the visible range.

Being a hypothetical substance, pure ice is defined as ice in which no scattering takes place under any circumstances. This definition makes the absorption of pure ice, and, hence, the imaginary part of the refraction index for pure ice, exceedingly difficult to measure. What we measure is the extinction coefficient $\sigma_t = \sigma_s + \sigma_a$, thus the problem is to find or grow ice in the real world which does not scatter light such that $\sigma_s = 0$. Such ice does not exist, but we can come close. Until the late nineties, the measurements reported by Grenfell and Perovich [1981] and by Perovich and Govoni [1991] have been used as the main reference for the imaginary part of the refractive index for pure ice n''_{ice}. They carefully grew a block of ice from a tank of filtered deionised water. They grew it from the bottom up to prevent air bubble inclusions thus coming as close to pure ice as possible in a laboratory.

Surprisingly, it seems to be the case that even purer ice than what they were able

λ [nm]	n'_{ice}	n''_{ice}
375	1.3222	$2.42 \cdot 10^{-11}$
400	1.3194	$2.37 \cdot 10^{-11}$
25	1.3174	3.52
50	1.3157	9.24
75	1.3143	$2.38 \cdot 10^{-10}$
500	1.3130	$5.89 \cdot 10^{-10}$
25	1.3120	$1.24 \cdot 10^{-9}$
50	1.3110	2.29
75	1.3102	3.80
600	1.3094	$5.73 \cdot 10^{-9}$
25	1.3087	9.50
50	1.3080	$1.43 \cdot 10^{-8}$
75	1.3075	1.99
700	1.3069	$2.90 \cdot 10^{-8}$
25	1.3064	4.17
50	1.3058	5.87
775	1.3054	$9.37 \cdot 10^{-8}$

Table 16.1: *Spectral index of refraction for pure ice. The real part is from the compilation of Warren* [1984]*, the imaginary part is from the work of Warren et al.* [2006] ($375\,\text{nm} \leq \lambda \leq 600\,\text{nm}$) *and the work of Grenfell and Perovich* [1981] ($600\,\text{nm} \leq \lambda \leq 780\,\text{nm}$).

to grow in the laboratory exists in nature. The Antarctic Muon and Neutrino Detector Array collaboration (AMANDA) found even lower values of spectral absorption (as compared to those of the laboratory grown ice) for ice 800-1800 meters deep in the Antarctic Ice Sheet at the South Pole [Askebjer et al. 1997]. At these depths most of the scattering inclusions have been dissolved in the ice as clathrates. This indicates that some amount of scattering has been present in the laboratory measurements. New values for the absorption coefficient of pure ice $\sigma_{a,\text{ice}}$ have therefore been measured indirectly in clean untouched snow by Warren, Brandt, and Grenfell [2006]. These newly measured absorption coefficients match those measured by Grenfell and Perovich [1981] from the wavelength $\lambda = 600\,\text{nm}$ and upwards. Using some of the mentioned references, I have collected measured values for the refractive index of pure ice. They are compiled in table 16.1.

Ice, as found in nature, is not only a weakly absorbing material, it also contains many different types of scattering inclusions. Most ice contains air particles and sea ice also contains brine pockets. In addition, minerals and algae may be present in the ice.

When sea ice gets very cold salts start precipitating. At temperatures below −8.2 °C mirabilite crystals start forming. Hydrohalite crystals form when temperature gets below −22.9 °C. Both these precipitated salts actually scatter light [Light et al. 2004], but to avoid making the example more complicated than necessary, I have chosen not to include the scattering of these crystals in the model.

Air Inclusions

Particles of air are easily modeled. They have a spherical shape and can be described quite accurately by the complex index of refraction $n_{\text{air}} = 1.00$. The optical properties of air differ only marginally from those of vacuum. The difference is of no consequence when we consider small bubbles. Air bubbles give us means to compute the optical properties of clean fresh water ice.

The number density of air bubbles follows a power law distribution [Grenfell 1983; Light et al. 2003]:

$$N_{\text{air}}(r) = N_{*,\text{air}} r^{-\alpha_{\text{air}}} \quad , \tag{16.1}$$

where α_{air} depends on freezing (or melting) conditions and the age of the ice. $N_{*,\text{air}}$ is a constant which is determined by the relationship between number density and volume fraction (10.4). In rapidly growing young sea ice going through its initial formation stage, Grenfell [1983] found $\alpha_{\text{air}} = 1.24$ and rather large bubbles with $r_{\text{min,air}} = 0.1\,\text{mm}$ and $r_{\text{max,air}} = 2\,\text{mm}$. If these limits are used in the relation between volume fraction and number density (10.4), it follows that

$$N_{*,\text{air}} = v_{\text{air}} \left(\frac{4\pi}{3} \int_{0.1}^{2} r^3 r^{-1.24}\,\mathrm{d}r \right)^{-1} = v_{\text{air}}/10.3 \quad .$$

In a section of interior first-year sea ice, Light et al. [2003] found $\alpha_{\text{air}} = 1.5$ and smaller bubbles with $r_{\text{min,air}} = 0.004\,\text{mm}$ and $r_{\text{max,air}} = 0.07\,\text{mm}$. All these bubbles were found inside brine pockets or tubes which means that their scattering cross sections should be computed with brine as host medium. Later Light et al. [2004] also classified previously unidentified features of their ice sample as air bubbles embedded directly in the ice. These bubbles are comparable to the ones measured by Grenfell [1983]. This means that air inclusions are modeled as two different types of particles: Those immersed in brine, which are called *active bubbles* and they follow the power law with exponent $\alpha_{\text{air},a} = 1.5$, and those embedded directly in the ice, which are called *inactive bubbles* and which follow the power law with exponent $\alpha_{\text{air},i} = 1.24$. The smaller active bubbles account for 8% of the total air volume while the larger inactive bubbles account for the remaining 92% of the total air volume.

Brine Inclusions

The refractive index of brine has already been described in the previous chapter, but brine particles in ice have temperatures below 0 ℃, so we need a different formula. Typical sea water freezes when temperatures get below −2 ℃. Then brine and ice is a phase system. If we assume that the system is always in phase equilibrium, there is a specific relation between the temperature and the salinity of the brine. This means that we can find a formula for the refractive index of ice that only depends on temperature.

An expression for the real part of the refractive index of brine n'_{brine} with temperatures in the interval $T \in [-32\,℃, -2\,℃]$ has been found by Maykut and Light [1995]. The expression they propose fit their measurements well, but it is based on the Lorentz-Lorenz relation (cf. Equation 8.4) and it requires mass-weighted molar refractivities for the principal constituents of standard seawater in freezing equilibrium. The mass concentrations necessary for reproduction of their expression are available as tabulated data from Richardson [1976], and the molar refractivity of pure water as an empirical function of wavelength should be retrieved from Schiebener et al. [1990]. The remaining refractivities are reproduced by Maykut and Light [1995] from data measured by Stelson [1990] and are considered independent of wavelength. All refractivities are considered independent of temperature (in the considered temperature interval). Finally the density of the brine is needed and, based on their measurements, Maykut and Light give a few different options (one of which is based on data measured by Thompson and Nelson [1956]) for determining this quantity as a function of temperature. For many applications, this expression is not practical.

As a simpler alternative I have chosen to extrapolate the empirical formula (15.1) by Quan and Fry [1995]. From the phase equilibrium table by Assur [1958] and that of Richardson [1976], one obtains the relationship between temperature and salinity of brine in freezing sea ice. Mirabilite crystals start precipitating in the brine at a temperature of −8.2 ℃ [Maykut and Light 1995; Light et al. 2004]. This causes large changes in the brine chemistry. Therefore we need a two-piece fit to the phase equilibrium data. Using a parabolic, least squares, two-piece fit with the empirical formula by Quan and Fry [1995], we obtain the following empirical formula

$$n'_{\text{brine}}(T,\lambda) = G_1(T) + \frac{G_2(T)}{\lambda} - \frac{4382}{\lambda^2} + \frac{1.1455 \cdot 10^6}{\lambda^3} \quad , \tag{16.2}$$

where λ is measured in nm and $G_i(T)$, $i = 1, 2$, have the form

$$G_i(T) = \alpha_0 - \alpha_1 T - \alpha_2 T^2 \quad .$$

The coefficients are given in Table 16.2. To reduce the number of coefficients, the terms involving temperature power three and four have been removed and

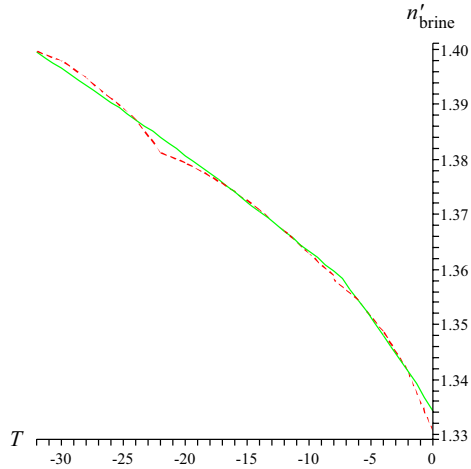


Figure 16.1: *The dependency of the refractive index for freezing brine on the temperature. Comparison of the new extrapolation of Quan and Fry's* [1995] *formula (the green, solid curve) to the not so practical fit of Maykut and Light* [1995] *(the red, dashed curve).*

the remaining coefficients have been corrected to make the curves meet at the breaking point $T = -8.2\,°\text{C}$. While being less detailed, since we only incorporate one of the many singularities along the curve, the resulting formula shows surprisingly good agreement with the much more complicated fit by Maykut and Light [1995]. See Figure 16.1.

The presented formula for the real part of the refractive index of freezing brine was found using the relation between temperature and salinity which is given by the phase equilibrium curves for a system of ice and brine. This relation was simply inserted in the empirical formula by Quan and Fry [1995], and we found

G_i	T [°C]	α_0	α_1	α_2
G_1	$[-8.2, -2]$	1.3144	$3.3344 \cdot 10^{-3}$	$5.2371 \cdot 10^{-5}$
G_1	$[-32, -8.2]$	1.3209	$2.2936 \cdot 10^{-3}$	$2.3099 \cdot 10^{-5}$
G_2	$[-8.2, -2]$	15.889	0.22059	$4.5478 \cdot 10^{-3}$
G_2	$[-32, -8.2]$	16.310	0.15053	$2.2718 \cdot 10^{-3}$

Table 16.2: *Coefficients for the empirical formula finding the real part of the refractive index* $n'_{\text{brine}}(T, \lambda)$ *of brine in freezing sea ice.*

good agreement with the fit by Maykut and Light [1995]. This suggests that the formula by Quan and Fry is valid for wider range of temperatures than originally assumed. The phase equilibrium curves are valid when the temperature of the brine is below its freezing point (c. −2 °C). Hence, there is a short interval of temperatures $-2\,^\circ\mathrm{C} \leq T \leq 0\,^\circ\mathrm{C}$ where we do not know if the formula is valid. However, since it is the same formula used both for temperatures below −2 °C and for temperatures above 0 °C, it is quite probable that the formula of Quan and Fry is valid in the entire range of temperatures from −32 °C to 30 °C.

As mentioned in Section 15.1, the imaginary part of the refractive index of brine n''_{brine} exhibits a weak dependency on temperature and a slight dependency on salinity in the visible range. It is, however, difficult to say if the correction formula by Pegau et al. [1997] is also valid for freezing brine. Therefore we will neglect these dependencies and simply use the imaginary part of the refractive index for brine at temperature $T = 0\,^\circ\mathrm{C}$. This concludes our model for the index of refraction of freezing brine. Let us find the shapes and number density distributions of brine particles in ice.

When water freeze in the sea, brine is trapped between vertical platelets of pure ice [Grenfell 1983]. If the freeze continues (below −2 °C), the initial platelet formation closes off and becomes scattering inclusions formed as cylindrical brine tubes, smaller ellipsoidal brine pockets, and very small spherical brine bubbles. As the air particles, the number density of the brine particles also follows a power law distribution. The distribution is [Light et al. 2003]

$$N_{\text{brine}}(\ell) = N_{*,\text{brine}}\, \ell^{-1.96} \quad , \tag{16.3}$$

where ℓ is the length of the non-spherical brine inclusions measured in mm. To fully describe the size of the brine inclusions, an empirical power law has also been found for the length-to-diameter aspect ratio γ of the brine inclusions [Light et al. 2003]:

$$\gamma(\ell) = \begin{cases} 1 & \text{for } \ell \leq 0.03\,\text{mm} \\ \gamma_*\, \ell^{\,0.67} & \text{for } \ell > 0.03\,\text{mm} \end{cases} \quad .$$

As is revealed by this relation, brine inclusions are considered spherical when $\ell \leq 0.03\,\text{mm}$. Brine inclusions of the size $0.03\,\text{mm} < \ell < 0.5\,\text{mm}$ are called *pockets* and are modelled as prolate ellipsoids. The remaining Brine inclusions, $\ell \geq 0.5\,\text{mm}$, are called *tubes* and are modeled as right circular cylinders [Light et al. 2004]. The size limits in the sample considered by Light et al. [2003; 2004] are $\ell_{\min} = 0.01\,\text{mm}$ and $\ell_{\max} = 14.6\,\text{mm}$. Brine bubbles account for 2% of the total brine volume, pockets for 6%, and tubes for the remaining 92%. The number densities of the different brine inclusions should be found using the volume fraction v_{brine} that would be found in ice of temperature $T = -15\,^\circ\mathrm{C}$ (the relation between ice temperature and the volume fraction of brine is described

in Section 16.2). At the temperature $T = -15\,°\text{C}$, we have $\gamma_* = 10.3$. Correction to obtain number densities for the true temperature is done by a proportional scale of bubbles and pockets, and a proportional scale of γ_* for the tubes. The proportionality constant is the relative change in the volume fraction of brine $v_{\text{brine}}(T)/v_{\text{brine}}(-15\,°\text{C})$.

To account for brine pockets and brine tubes in the Lorenz-Mie calculations, we use the volume-to-area equivalent spheres described in Section 9.3. To find the volume-to-area equivalent spheres, we need the volume and surface area of prolate ellipsoids and right circular cylinders. In the following, we denote brine pockets by the subscript bp and brine tubes by the subscript bt. The surface area of a prolate ellipsoid (or spheroid) is

$$A_{\text{bp}} = \frac{\pi}{2}\frac{\ell^2}{\gamma^2}\left(1 + \gamma\frac{\sin^{-1}\varepsilon}{\varepsilon}\right) \quad ,$$

where ε is the ellipse eccentricity defined by

$$\varepsilon = \sqrt{1 - 1/\gamma^2} \quad .$$

The volume is

$$V_{\text{bp}} = \frac{\pi}{6}\frac{\ell^3}{\gamma^2}$$

which means that the radius $r_{\text{eq,bp}}$ of the volume-to-area equivalent spheres is (9.25)

$$r_{\text{eq,bp}} = \frac{\ell\,\varepsilon}{\varepsilon + \gamma\sin^{-1}\varepsilon} \quad .$$

For right circular cylinders we find

$$\begin{aligned} A_{\text{bt}} &= \frac{\pi}{2}\frac{\ell^2}{\gamma^2}(1 + 2\gamma) \\ V_{\text{bt}} &= \frac{\pi}{4}\frac{\ell^3}{\gamma^2} \\ r_{\text{eq,bt}} &= \frac{3\,\ell}{2 + 4\gamma} \quad . \end{aligned}$$

This gives us all the information we need to find the relative number density of the volume-to-area equivalent spheres for both the brine pockets and the brine tubes. Using Equation 9.26, we have

$$\bar{N}_{\text{eq}}(r_{\text{eq}}) = \frac{N_{\text{eq}}(r_{\text{eq}})}{N_*} = \frac{N_{\text{brine}}(r_{\text{eq}})}{N_{*,\text{brine}}}\frac{3V}{4\pi r_{\text{eq}}^3} = r_{\text{eq}}^{-1.96}\frac{3V}{4\pi r_{\text{eq}}^3} \quad .$$

The relation between volume fraction and number density (10.4) is used to find the unknown factors $N_{*,\text{bp}}$ and $N_{*,\text{bt}}$ in the distributions. We have

$$N_{*,i} = v_i\left(\int_{\ell_{\min}}^{\ell_{\max}} (r_{\text{eq},i}(\ell))^{-1.96} V_i(\ell)\,\mathrm{d}\ell\right)^{-1} \quad ,$$

Property	Pure ice (host)	Air	Brine
n'	Table 16.1	1.00	Equation 16.2
n''	Table 16.1	0.00	Table 15.2
r		[0.1 mm, 2.0 mm]	[0 mm, 14.6 mm]
$N(r)$		Equation 16.1	Equation 16.3

Table 16.3: *A summary of the microscopic properties of sea ice.*

where i is either bp or bt. The microscopic properties of sea ice are summarised in Table 16.3.

16.2 Appearance Model

The information provided in the previous section makes us able to create a detailed appearance model for ice. The parameters are

- Ice temperature T
- Volume fraction of air in the ice v_{air}
- Volume fraction of brine in the ice v_{brine} .

Unfortunately the volume fractions of air and brine in the ice are not very useful parameters. They are much more difficult to measure than, for example, volume fractions of minerals and algae. The problem is that they are very sensitive to temperature changes. Let us see if we can exploit this dependency on temperature to get a better set of parameters.

Cox and Weeks [1983] have found a relation between three common physical parameters and the volume fractions of air and brine in ice. The three parameters are ice density ρ measured in g/mL, ice temperature T in ℃, and ice salinity S

F_i	α_0	α_1	α_2	α_3
$-2\,°\text{C} \leq T < 0\,°\text{C}$				
F_1	$-4.221 \cdot 10^{-2}$	$-1.8407 \cdot 10^{1}$	$5.4802 \cdot 10^{-1}$	$2.4154 \cdot 10^{-1}$
F_2	$9.0312 \cdot 10^{-2}$	$-1.6111 \cdot 10^{-2}$	$1.2291 \cdot 10^{-4}$	$1.3606 \cdot 10^{-4}$
$-22.9\,°\text{C} \leq T < -2\,°\text{C}$				
F_1	-4.732	$-2.245 \cdot 10^{1}$	$-6.397 \cdot 10^{-1}$	$-1.074 \cdot 10^{-2}$
F_2	$8.903 \cdot 10^{-2}$	$-1.763 \cdot 10^{-2}$	$-5.330 \cdot 10^{-4}$	$-8.801 \cdot 10^{-6}$
$-32\,°\text{C} < T \leq -22.9\,°\text{C}$				
F_1	$9.899 \cdot 10^{3}$	$1.309 \cdot 10^{3}$	$5.527 \cdot 10^{1}$	$7.160 \cdot 10^{-1}$
F_2	8.547	1.089	$4.518 \cdot 10^{-2}$	$5.819 \cdot 10^{-4}$

Table 16.4: *Coefficients for the empirical formulae (16.4–16.5) of Cox and Weeks [1983]. The coefficient in the temperature range $-32\,℃ \leq T \leq -2\,℃$ were reported in the original reference. The additional set of coefficients for the temperature range $-2\,℃ \leq T \leq 0\,℃$ was reported by Leppäranta and Manninen [1988].*

in parts per thousand ‰. The relations are

$$v_{\text{air}} = 1 - \frac{\rho}{\rho_{\text{pure}}} + \rho S \frac{F_2(T)}{F_1(T)} \tag{16.4}$$

$$v_{\text{brine}} = \frac{\rho S}{F_1(T)} \tag{16.5}$$

$$\rho_{\text{pure}} = 0.917 - 1.403 \cdot 10^{-4}\, T \ , \tag{16.6}$$

where F_i, $i = 1, 2$, have the form

$$F_i(T) = \alpha_0 + \alpha_1 T + \alpha_2 T^2 + \alpha_3 T^3 \tag{16.7}$$

and the coefficients are given in Table 16.4. Note that F_1 is measured in g/mL, while F_2 is dimensionless.

These empirical relations improve our appearance model, because now the microscopic properties follow from the parameters

- Ice temperature T
- Ice salinity S
- Ice density ρ .

Through variation of these three parameters, we are able to compute the macroscopic optical properties of many different types of ice. If we assume that mineral and algal particles appear in ice in the same way as they appear in water, we can add another two parameters to the appearance model using the microscopic properties described in Section 15.1. The two extra parameters are

- Volume fraction of minerals v_{mineral}
- Volume fraction of algae v_{alga} .

These five parameters have an intuitive physical meaning. Therefore we are able to specify sensible intervals for them to live in, and we are able to explain their effect on the appearance of the ice, qualitatively, before starting a rendering. This gives a modeler or animater, who needs to render ice, a good chance of setting sensible values instead of spending too much time doing manual adjustment of optical properties.

Let us try to specify sensible intervals for some of the parameters. The density of pure ice (16.6) is close to 0.917 g/mL and it increases slightly with decreasing temperature. This means that the density of an iceberg typically is in the range $0.86\,\text{g/mL} < \rho < 0.93\,\text{g/mL}$, where ice with $\rho = 0.86\,\text{g/mL}$ is warm and very bubbly, while ice with $\rho = 0.93\,\text{g/mL}$ is cold, very saline, and almost bubble-free. The ice salinity is in the range $0‰ < S < 12‰$, where $S = 0$ denotes fresh water ice. Last we have temperature which is typically in the range $-32\,℃ < T < 0\,℃$. Sea water will, however, not start freezing until the temperature falls below $-2\,℃$.

Potential Expansion of the Model

Considering an iceberg, non of these physical parameters would be constant throughout the bulk medium. They would rather be constant in layers and change as we move from the top, say $z = 0$, towards the bottom $z_b > 0$ of the berg. If we consider a young first-year ice sheet before the onset of the melting period, the ice temperature increases linearly with z from the air temperature to around $-2\,℃$ (at the bottom). Ice salinity, on the other hand, is high in the top 10 cm of the sheet, but decreases until around 10 cm from the bottom where it starts increasing again. The relationship between density and z is harder to determine. The density ρ should decrease as the ice gets warmer with increasing z, but on the other hand ρ also increases as the pressure of the above ice dissolves air inclusions into clathrates with increasing z. These depth profiles for icebergs could be captured using the geometrical approach described in Part III.

Pure ice Compacted ice White ice

Figure 16.2: *The Stanford dragon model rendered using different types of ice. The dragon is 50 meters long and at this size it shows the effect of absorption by the ice. This absorption is the reason for the blue light transmitted through the pure ice, the cold blue light of the compacted ice, and the deep blue light in the shadow regions of the white ice dragon.*

16.3 Results

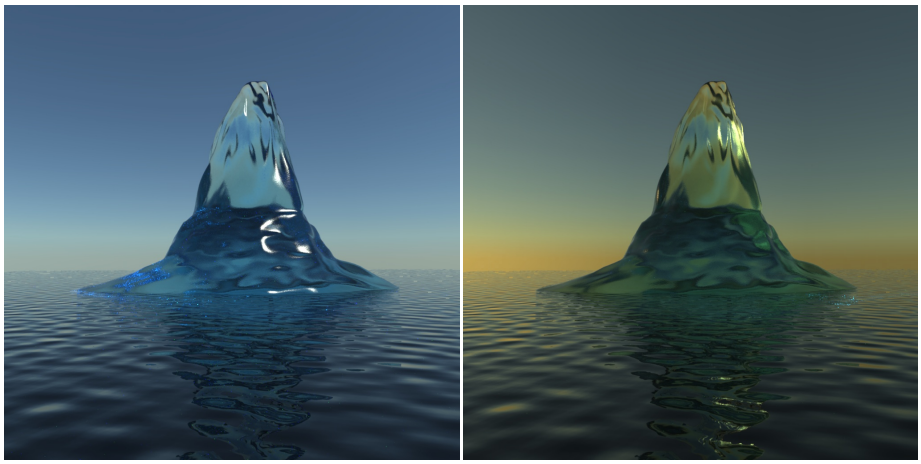
Since our appearance model is closely coupled to the physical properties of the ice, it is easy to take measured physical properties of ice reported in the literature and convert them to the appearance of the ice. As examples, we have found optical properties of ice resulting from compacted snow and white first-year ice. Renderings of these two types of ice as well as pure ice for comparison, are presented in Figure 16.2. The contents of brine and air in compacted ice are $v_{\text{brine}} = 8.0 \cdot 10^{-4}$ and $v_{\text{air}} = 2.3 \cdot 10^{-4}$. The parameters used for the white first-year ice are temperature $T = -15\,^{\circ}\text{C}$, salinity $S = 4.7\,‰$, and density $\rho = 0.921\,\text{g/mL}$. These are the parameters measured in first-year sea-ice samples by Light et al. [2004].

Light et al. [2004] also measured a reduced scattering coefficient of $\sigma_s' \approx 8$ for the same first-year sea ice samples. This provides an opportunity for testing our theory as well as the choice of using volume-to-area equivalent spheres (see Sec. 9.3). If we go through the computations described in Part II and use volume equivalent spheres, we arrive at the scattering coefficient $\sigma_s' \approx 6.4$. If we use volume-to-area equivalent spheres, we get the much more correct value $\sigma_s' \approx 8.1$.

As another example, let us consider the peculiar bottle green icebergs which are sometimes encountered in Antarctic regions. This type of iceberg is almost free of brine and air inclusions [Dieckmann et al. 1987; Warren et al. 1993]. They do, however, contain some gray scattering inclusions that we model as minerals. To account for these minerals in the ice, we change the exponent used

with the number density distribution in the ocean water case to $\alpha = 1.5$. We then include a tiny fraction of brine and a very small fraction of air, as well as a volume fraction of minerals which is $v_{\text{mineral}} = 1.5 \cdot 10^{-7}$. A theory for the bottle green sometimes observed in these very clean icebergs, is that their colour depends on the light incident through the atmosphere [Lee 1990]. Therefore the ice will look clean and very transmissive during the day, but when sun begins to set, and the spectrum of the incident light shifts towards red wavelengths, the minerals in the iceberg shift the spectrum such that the ice gets a dark green appearance. The rendered bottle green iceberg shown in Figure 16.3 illustrates the very different appearance of this type of iceberg at two different times of the day. The images show that Lee's theory could be right. Other green icebergs have been observed with a high algal content. If we include the same algal and mineral contents as the contents found in the North Sea (see Table 15.5), the iceberg gets the green colour shown in Figure 16.4. Unlike the iceberg in Figure 16.3, the iceberg in Figure 16.4 exhibits a green colour at all times during the day.

In this chapter we have seen that it is possible to make an appearance model which follows the physical conditions of the material. These conditions are naturally coupled to the conditions of the surroundings. Such a coupling is useful in two distinct ways. One application is to predict the appearance of a material under different physical conditions. Another application is to judge the conditions of the surroundings by looking at a material. In the next chapter we will look at milk, which is an interesting example because macroscopic optical properties have been measured in graphics using camera technology. This provides a good opportunity to see if the theory works and if we are able to make conclusions about the material using the measurements.



10am 7pm

Figure 16.3: *A blue iceberg at noon turns green in the evening. These bottle green icebergs are one of nature's peculiarities. We simulated the properties of the green iceberg shown in these images by including a small amount of minerals and only very little air and brine in the ice. The atmospheric lighting model uses Rayleigh scattering to obtain the spectral radiance for the skylight and the sun.*

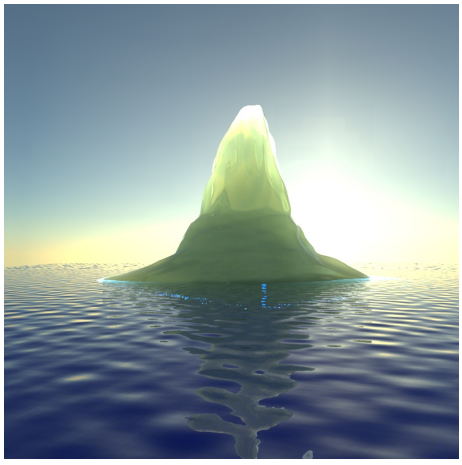


Figure 16.4: *An iceberg with low air and brine contents, but high algal content. This iceberg is green at all times during the day.*

Chapter 17

Milk

Even small amounts of milk products can improve the health of people who are forced to subsist on the edge of starvation.

Marvin Harris, from *Cows, Pigs, Wars, and Witches*

Milk consists roughly of an emulsion of milkfat globules; a colloidal suspension of protein particles; and lactose, soluble proteins, minerals, vitamins, acids, enzymes, and other components dissolved in water [Goff and Hill 1993]. About 80% of the protein in milk is casein protein. Most of this casein, about 95% [Fox and McSweeney 1998], exists in colloidal particles known as casein micelles. From an optical point of view, milk can then be treated as two different types of spherical particles, namely fat globules and casein micelles, suspended in a host medium with almost the same optical properties as pure water. The absorption spectrum of the host medium needs to be adjusted because of dissolved vitamin B2 (riboflavin) which exhibits absorption in the visible range of the spectrum. In fact riboflavin is also fluorescent, but we will not take that into account. What we use is, in other words, a simplified model of milk, but it should be sufficient for considering the appearance of milk.

17.1 Particle Composition

The host medium is water in which many different components are dissolved. For the real part of the refractive index of the milk host, we use the refractive

index for pure fresh water ($S = 0\,‰$ in Equation 15.1).

To find the imaginary part of the refractive index of the milk host, we make a correction for the imaginary part of the refractive index of pure water. The dissolved component exhibiting the most significant absorption in the visible range is vitamin B2 (riboflavin). Spectral data for the absorption of vitamin B2 are available in the PhotochemCAD application[1] [Du et al. 1998]. The absorption coefficient is not measured directly, instead the absorbance D is measured. The absorption coefficient is calculated from the absorbance using a molar absorption coefficient ε for some wavelength. A molar absorption coefficient of riboflavin is $\varepsilon(266.5\,\text{nm}) = 3.3 \cdot 10^6\,\text{M}^{-1}\text{m}^{-1}$ [Koziol 1966]. The following formula finds the remaining molar absorption coefficients for riboflavin using the absorbance data:

$$\varepsilon(\lambda) = \frac{\varepsilon(266.5\,\text{nm})}{D(266.5\,\text{nm})} \ln(10) D(\lambda) \ .$$

The natural content in milk of riboflavin is 17 mg per 100 g milk. Using the molar mass of riboflavin which is 376.3682 g/mol, we find that the natural concentration of riboflavin in milk is

$$c = \frac{17\,\text{mg}}{376.3682\,\text{g/mol}} \Big/ \frac{100\,\text{g}}{1.03\,\text{g/mL}} = 4.65 \cdot 10^{-4}\,\text{mol/L} \ .$$

By multiplication of the molar absorption coefficient with this concentration, we obtain the absorption coefficient of riboflavin which is converted to the imaginary part of a refractive index (4.43) and added to the imaginary part of the refractive index for pure water. The result, n''_{milk}, is presented in Table 17.1.

The Fat Inclusion

Walstra and Jenness [1984] have found experimentally that the real part of the refractive index of milk fat approximately follows the function

$$n'_{\text{fat}}(\lambda, T) = \sqrt{\frac{(b(T)+2)\lambda^2 - 0.03}{(b(T)-1)\lambda^2 - 0.03}} \ , \tag{17.1}$$

where wavelength is measured in μm and b is found using a measurement of the refractive index at the temperature T. We use measurements by Michalski et al. [2001] since they also present the wavelength dependent imaginary part of the refractive index. They find $n'_{\text{fat}}(0.589\,\mu\text{m}, 20\,℃) = 1.461$ which gives $b(20\,℃) = 3.73$. This corresponds well to the $b(40\,℃) = 3.77$ reported by Walstra and Jenness [1984]. Table 17.1 includes the imaginary part of the

[1] http://omlc.ogi.edu/spectra/PhotochemCAD/abs_html/riboflavin.html

λ [nm]	n''_{milk}	n''_{fat}
375	$2.93 \cdot 10^{-7}$	$4.0 \cdot 10^{-6}$
400	$2.60 \cdot 10^{-7}$	$6.4 \cdot 10^{-6}$
25	3.36	8.6
50	4.10	$1.1 \cdot 10^{-5}$
75	3.33	1.1
500	$1.08 \cdot 10^{-7}$	$1.0 \cdot 10^{-5}$
25	$5.64 \cdot 10^{-8}$	$4.7 \cdot 10^{-6}$
50	6.02	4.6
75	7.91	4.7
600	$7.95 \cdot 10^{-8}$	$4.9 \cdot 10^{-6}$
25	8.58	5.0
50	9.32	5.0
75	7.37	5.1
700	$1.14 \cdot 10^{-7}$	$5.2 \cdot 10^{-6}$
25	1.33	5.2
50	2.20	5.2
775	$2.35 \cdot 10^{-7}$	$5.2 \cdot 10^{-6}$

Table 17.1: *Imaginary part of the refractive index for the milk host n''_{milk} and milk fat n''_{fat}. The milk host spectrum is a correction of the spectrum for pure water according to the content of dissolved vitamin B2 in the milk. The milk fat spectrum is from Michalski et al.* [2001].

refractive index for milk fat n''_{fat} as I read it from the curve reported by Michalski et al. [2001].

The volume frequency of the fat globules follows a log-normal distribution [Walstra 1975] (cf. Equation 10.2). The mean of the volume-to-area equivalent sphere radii $r_{\text{va,fat}}$ of the fat globules change depending on the volume fraction of the globules in the milk. By a least-squares, two-piece fit to measured data reported by Olson et al. [2004], I have found a functional expression describing this relationship:

$$r_{43,\text{fat}} = \begin{cases} -0.2528\, w_f^2 + 1.419\, w_f & \text{for } w_f < 2.0 \\ 1.456\, w_f^{0.36} & \text{otherwise} \end{cases} , \tag{17.2}$$

where $r_{43,\text{fat}}$ is measured in μm. The relationship between $r_{43,\text{fat}}$ and $r_{\text{va,fat}}$ is [Walstra 1975]

$$r_{\text{va,fat}} = r_{43,\text{fat}} / (c_{v,\text{fat}}^2 + 1) \ . \tag{17.3}$$

The radius $r_{43,\text{fat}}$ is used since it can be estimated empirically with good ac-

curacy [Walstra 1975]. The coefficient of variation $c_{v,\mathrm{fat}}$ is usually between 0.4 and 1.2 in normal milk. Reasonable limits for the range of fat globule radii are $r_{\mathrm{min,fat}} = 0.005\,\mu\mathrm{m}$ and $r_{\mathrm{max,fat}} = 10\,\mu\mathrm{m}$.

The Protein inclusion

The refractive index of casein micelles is not readily available in the literature. For comparison to goat's milk it has been determined to be the following for cow's milk [Attaie and Richtert 2000]:

$$n_{\mathrm{casein}} = 1.503 \ .$$

This value is assumed to be constant in the visible range and absorption of the casein micelles is neglected.

Structure and size distribution of casein micelles is still being disputed in the literature. Recent research on the matter is discussed by Gebhardt et al. [2006]. Most investigations are based on either light scattering or electron microscopy. Light scattering approaches find micelles of large average size while electron microscopy report a large number of very small casein particles in addition to the larger micelles. Sometimes these very small particles are excluded from the reported size distribution since they are regarded to represent non-micellar casein or single sub-micelles. No matter what we call these very small particles, they scatter light as do the larger aggregates and therefore should be included in the size distribution employed for the Lorenz-Mie calculations.

A size distribution based on electron microscopy, which includes the single sub-micelles in the distribution, was reported by Schmidt et al. [1973]. They found $r_{\mathrm{va,casein}} = 43\,\mathrm{nm}$ and showed that a log-normal distribution (10.2) of $r/(r_{\mathrm{max,casein}} - r)$ is a good fit of the measured volume frequency distribution. The limits for the casein micelle radii are $r_{\mathrm{min,casein}} = 0\,\mathrm{nm}$ and $r_{\mathrm{max,casein}} = 150\,\mathrm{nm}$.

The microscopic properties of milk are summarised in Table 17.2.

17.2 Appearance Model

To model the concentration of fat and protein we use wt.-% (g per 100 g milk), since this value is used on contents declarations on the side of milk cartons. In the remainder of this chapter we let w_f and w_p denote the wt.-% of fat

Property	Milk host	Fat globules	Casein micelles
n'	Equation 15.1	Equation 17.1	1.503
n''	Table 17.1	Table 17.1	0.00
r		$[0.005\,\mu\text{m}, 10\,\mu\text{m}]$	$[0\,\text{nm}, 150\,\text{nm}]$
$N(r)$		Equation 10.2	Equation 10.2

Table 17.2: *A summary of the microscopic properties of cow's milk.*

ρ_{fat}	ρ_{protein}	ρ_{milk}
$1.11\,\text{g/mL}$	$0.915\,\text{g/mL}$	$1.03\,\text{g/mL}$

Table 17.3: *Densities for computing of milk fat and casein volume fractions using weight percents. Measured by Walstra and Jenness* [1984] *at* 20 ℃.

and protein respectively. To translate wt.-% into volume fractions, we use the densities given by Walstra and Jenness [1984]. They are summarised in Table 17.3. Casein micelles make up about 76% of the protein volume fraction [Fox and McSweeney 1998]. This means that

$$v_{\text{fat}} = \frac{w_f/\rho_{\text{fat}}}{100\,\text{g}/\rho_{\text{milk}}} \quad \text{and} \quad v_{\text{casein}} = 0.76\frac{w_p/\rho_{\text{protein}}}{100\,\text{g}/\rho_{\text{milk}}} \ .$$

This simple translation from fat and protein contents to volume fractions of the particle inclusions in the milk means that we have an appearance model with the following parameters:

- Fat content w_f
- Protein content w_p .

These two parameters are all we need to model most types of milk.

Since the macroscopic optical properties of milk have previously been measured in graphics [Jensen et al. 2001; Narasimhan et al. 2006], we have the opportunity to analyse the properties of different milk samples using the measurements. To construct a geometrical representation of the milk appearance model, we find a set of control points (using Maple) that fit the computations from the

microscopic properties of the milk to the macroscopic optical properties of the milk. We include the following variables in our geometric representation:

- Fat content w_f
- Protein content w_p
- Mean size of the fat globules $r_{\text{va,fat}}$
- Coefficient of variation for the fat globules $c_{v,\text{fat}}$
- RGB colour representations for the bulk absorption coefficient σ_a
- RGB colour representations for the bulk scattering coefficient σ_s
- RGB colour representations for the ensemble asymmetry parameter g .

These constitute the axes in our multidimensional shape. The latter three are the macroscopic optical properties that have been measured using camera technology [Jensen et al. 2001; Narasimhan et al. 2006] (each of these macroscopic optical properties has three axes associated with them, one for each colour component).

Note that the constraint given by Equations 17.2 and 17.3 binds the mean size of the fat globules $r_{\text{va,fat}}$ when the fat content w_f is given. Nevertheless, it turns out that it is useful to include $r_{\text{va,fat}}$ in the model for the analysis.

Finally, it is sometimes practical to have a functional expression that return the macroscopic optical properties in RGB given a few simple input parameters. Appendix C is a fit of the Lorenz-Mie computations which takes only the fat and protein contents as input.

17.3 Results

To analyse the measured macroscopic optical properties, we take a slice in our geometrical representation using the values $c_{v,\text{fat}} = 0.6$ and $w_p = 3\,\text{wt.-\%}$ which are commonly found in milk. Then we do an orthogonal projection of the $r_{\text{va,fat}}$ axis to obtain the relation between the macroscopic optical properties and the fat content of the milk. The resulting relation is plotted in Figure 17.1 along with the measurements by Jensen et al. [2001] and by Narasimhan et al. [2006]. We use the reduced scattering coefficient in the figure because this is the quantity that was measured by Jensen et al. [2001]. The measurements have been placed

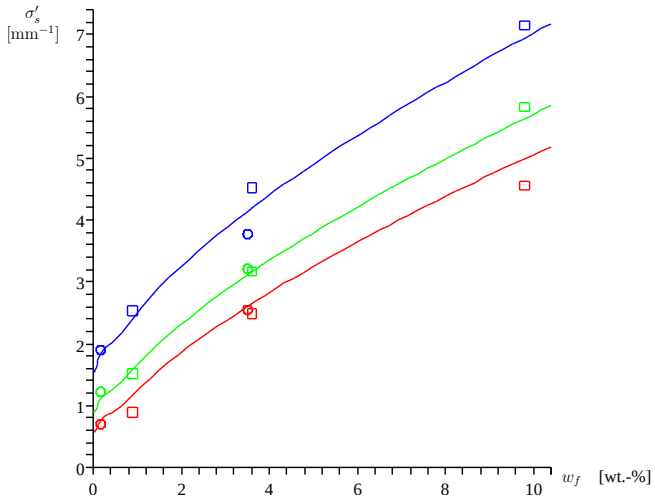


Figure 17.1: *Reduced scattering coefficients $\sigma'_s = (1-g)\sigma_s$ measured by Jensen et al. [2001] (circles) and Narasimhan et al. [2006] (squares) and plotted against our appearance model for milk as a function of fat content w_f.*

where they best fit the curves. The estimated fat contents for the measurements by Jensen et al. [2001], which are given by the position of the measurements along the horizontal axis in Figure 17.1, are most probably the fat contents of the original milk samples. For the measurements by Narasimhan et al. [2006] the fat contents have been overestimated (except for the lowfat case).

To analyse the reason for the overestimation, we pick a slice in our geometrical representation using $w_p = 3\,\text{wt.-\%}$. If we insert the measurements and an expected value for the fat content, we observe that the mean particle size and coefficient of variation for the fat globules, $r_{\text{va,fat}}$ and $c_{v,\text{fat}}$, become overestimated. This means that our computations match the measurements by Narasimhan et al. [2006] if we use the expected value for the fat content, but a larger $r_{\text{va,fat}}$ and $c_{v,\text{fat}}$ than what are commonly found in milk. An increase of these two parameters corresponds exactly to the effect of milk dilution [Walstra 1975]. Thus we have found by analysis that the milk measured by Narasimhan et al. [2006] was diluted, and indeed it was (as is obvious from the title of the reference).

A short comment on the absorption in milk: Because milk is highly scattering, its absorption coefficient is very difficult to measure. Measured absorption coefficients are, however, still of the same order of magnitude as the predicted ones and they all exhibit the behaviour $\sigma_{a,R} < \sigma_{a,G} < \sigma_{a,B}$. If we had ignored the absorption in the host medium in the Lorenz-Mie calculations, the absorption coefficient would only depend on fat content and would under-estimate the ab-

sorption in the blue band by more than a factor 48 for skimmed milk, while the smallest error would be −96% in the red band.

As the climax of this thesis, Figure 17.2 shows several glasses containing water, vitamin B2, protein and fat in various combinations. These have all been rendered using the appearance model described in this chapter. The resulting images of milk capture important visual properties such as the red shadow in skimmed milk and the increasingly white appearance as the fat content increases. The images also show how the link between microscopic and macroscopic optical properties makes us able to visualise different components in a material independently. Knowing the visual significance of the different ingredients in a material is important both if we would like to design the appearance of the material or if we would like to interpret the appearance of the material. In the next chapter, which is also the final chapter, I will draw conclusions on the work presented and provide more perspective on the range of potential applications.

Figure 17.2: *Rendered images of the components in milk (top row) as well as mixed concentrations (bottom row). From top left to bottom right the glasses contain: Water, water and vitamin B2, water and protein, water and fat, skimmed milk, regular milk, and whole milk.*

Chapter 18

Conclusion

Beyond the horizon
Behind the sun
At the end of the rainbow
Life has only begun

Bob Dylan from *Modern Times*

The ice example (Chapter 16) was the very beginning of the work presented in this thesis. In fact, the original intension was to write a paper on iceberg rendering. In trying to find the correct index of refraction for ice, I learned that ice has a complex index of refraction. This was a surprise because I had only heard about complex indices of refraction in the context of metals. It was only then that I learned about the direct relationship between absorption and the imaginary part of the index of refraction. Led by curiosity, I wanted to find out how the law of refraction works with a complex index of refraction (Chapter 4). In this way I found out that there are quite a few problems associated with geometrical optics in absorbing media (Chapter 5).

Finding that the problems are negligible when we do not consider geometrical optics in small absorbing particles, I proceeded to investigate the scattering of particles in ice. Again the absorption of pure ice (the complex index of refraction) turned out to be causing problems. Pure ice is the weakly absorbing host of the air bubbles and brine pockets in icebergs. The scattering cross section turned out to be problematic for a particle in an absorbing host (Chapter 6). Moreover the Lorenz-Mie theory also had problems in dealing with an absorbing

host (Chapter 9). The first of these problems is still a problem. The second turned out to be solvable within the time frame of this thesis.

The next problem was to connect scattering cross sections to the optical properties used for realistic rendering (Chapters 6 and 10). When this link had been made I was ready to make both icebergs and milk and natural waters (Chapters 15–17). These first rough appearance models made me realise that one of the most important tasks in the specification of an appearance model is to arrive at sensible input parameters. This led to the specification of an appearance model for ice which only depends on temperature, salinity, and density of the ice (and perhaps some content of minerals and algae), and a model for milk which only depends on fat and protein contents. The specification of these appearance models was a rather cumbersome process which suggested that some framework should be available for handling the many parameters in the appearance of an object. This in turn led to the ideas presented in Part III.

Since volume rendering is a slow process, I eventually became curious about the connection between volume and surface rendering (Chapter 7). This led to the new derivation of Fick's law of diffusion, and I realised the significant limitations of diffusion-based rendering methods. Fast realistic rendering of translucent materials is still a significant challenge in computer graphics.

With this large body of theoretical work, the thesis became an investigation into the foundations of appearance modelling rather than an investigation of all the details in the perfect modelling and rendering of a specific material. Although it is unusual for a thesis to spread over such a large body of theory, I believe it is justified because the discipline of appearance modelling is dominated by detailed work on specific materials. There is nowhere else to find a connection from quantum electrodynamics to surface-based rendering techniques. This connection opens up for a better understanding of the influence of physical properties on the appearance of materials.

As the final remarks, let us discuss the potential applications of the presented connection between physics and appearance. The milk example (Chapter 17, in particular Figure 17.2) demonstrates that we are able to show the visual significance of the different components in a material. It also demonstrates that we are able to *predict* the appearance for various ratios between the contents. This makes a number of interesting applications possible:

- If you want to *design* materials with a specific appearance, the appearance model can help you choose the right components to obtain the desired appearance.
- If you want to *detect* whether a component is present in a material or not,

> the appearance model can help you visualise the material as it would look with and without the component.

In other words, synthesised images with a connection to the contents and the physical conditions of a material enables us to learn a lot about the reasons for the appearance of materials.

The techniques available in graphics for measuring macroscopic optical properties using camera technology enable interesting new applications in combination with the theory presented in this thesis. The comparison between computed and measured optical properties of milk (Figure 17.1) demonstrates that we are able to draw sensible conclusions about the original fat contents of the milk (especially if the measurements are not based on dilution). With future research and development, this could potentially turn into a new type of equipment which is able to measure the contents of materials using simple pictures.

Suppose we have an appearance model for a material and represent it as a multidimensional shape (Part III). Then geometric operations will be able to retrieve the appearance of the material under various circumstances (by picking slices). They will also be able to retrieve all the conclusions that we can draw about a material given, for example, measurements of its macroscopic optical properties (using projection). The concept of multidimensional shapes provides the perfect tool both for synthesising appearance and for analysing appearance. However, to construct the multidimensional shape, we first need an appearance model which connects appearance to the physical properties of the material. Such an appearance model is obtained by considering the composition of the material and the interaction of light and matter at a microscopic level (Part II). Once the appearance model is available, the material is rendered using a macroscopic theory of light (Part I). It is from this point of view that I consider light, matter, and geometry to be the cornerstones of appearance modelling.

Appendix A

On Geometrical Optics

And this particular camel, the result of millions of years of selective evolution to produce a creature that could count the grains of sand it was walking over, and close its nostrils at will, and survive under the broiling sun for many days without water, was called You Bastard.

And he was, in fact, the greatest mathematician in the world.

Terry Pratchett, from *Pyramids*

A.1 Second Order Wave Equations

As in the usual treatment of Maxwell's equations [Born and Wolf 1999, for example], take Faraday's law (4.31) divide by μ and apply the curl operator on both sides of the equality:

$$\nabla \times (\mu^{-1} \nabla \times \boldsymbol{E}_c) = i\omega \nabla \times \boldsymbol{H}_c \ .$$

Inserting from the first Maxwell equation (4.30) gives

$$\nabla \times (\mu^{-1} \nabla \times \boldsymbol{E}_c) = \omega^2 (\varepsilon + i\sigma/\omega) \boldsymbol{E}_c \ .$$

To move μ^{-1} outside the curl on the left-hand side, we use the identity from vector calculus: $\nabla \times u\boldsymbol{v} = u\,\nabla \times \boldsymbol{v} + \nabla u \times \boldsymbol{v}$. Thereby

$$\nabla \times (\nabla \times \boldsymbol{E}_c) + \mu\, \nabla \mu^{-1} \times (\nabla \times \boldsymbol{E}_c) = \omega^2 \mu (\varepsilon + i\sigma/\omega) \boldsymbol{E}_c = k_0^2 n^2 \boldsymbol{E}_c \ .$$

Or if we use $u \nabla u^{-1} = -\nabla \ln u$, another way to write it is

$$\nabla \times (\nabla \times \boldsymbol{E}_c) - \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_c) = k_0^2 n^2 \boldsymbol{E}_c \ .$$

Finally using $\nabla \times (\nabla \times) = \nabla(\nabla \cdot) - \nabla^2$, we find

$$\nabla^2 \boldsymbol{E}_c - \nabla(\nabla \cdot \boldsymbol{E}_c) + \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_c) = -k_0^2 n^2 \boldsymbol{E}_c \ . \tag{A.1}$$

Another identity from vector calculus is $\nabla \cdot u\boldsymbol{v} = u \nabla \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \nabla u$. This makes us able to write the third Maxwell equation (4.32) as

$$\varepsilon \nabla \cdot \boldsymbol{E}_c + \boldsymbol{E}_c \cdot \nabla \varepsilon = 0 \ .$$

Division by ε, using $u^{-1} \nabla u = \nabla \ln u$, and taking the gradient on both sides of the equation gives

$$\nabla(\nabla \cdot \boldsymbol{E}_c) = -\nabla(\boldsymbol{E}_c \cdot \nabla \ln \varepsilon) \ .$$

But then Equation A.1 reduces to

$$\nabla^2 \boldsymbol{E}_c + \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_c) + \nabla(\boldsymbol{E}_c \cdot \nabla \ln \varepsilon) = -k_0^2 n^2 \boldsymbol{E}_c \ ,$$

which is sometimes referred to as a *second order wave equation*.

Following a similar line of arguments, a formally identical equation is found for the magnetic vector:

$$\nabla^2 \boldsymbol{H}_c + \nabla \ln \varepsilon \times (\nabla \times \boldsymbol{H}_c) + \nabla(\boldsymbol{H}_c \cdot \nabla \ln \mu) = -k_0^2 n^2 \boldsymbol{H}_c \ .$$

The reason why these equations are called second order wave equations is that if ε and μ do not depend on position, the equations are the same as

$$\nabla^2 \boldsymbol{F} = n^2 \frac{\partial^2 \boldsymbol{F}}{\partial t^2} \ ,$$

which is an ordinary (or first order) wave equation, and $\boldsymbol{F}$ could be either $\boldsymbol{E}_c$ or $\boldsymbol{H}_c$ in a homogeneous medium.

A.2 The Eikonal Equation

In this appendix the eikonal equation (5.7) is derived by insertion of the wave function (5.5) in the second order wave equation (5.1) for the electric field. The second order wave equation was derived in Section A.1. Note that the time exponential cancels out upon insertion of the wave function as it appears in every term of the equation. Therefore we use the wave function (5.5) without the time exponential in this appendix. During the derivation, we employ the

following identities from vector calculus:

$$\begin{aligned} \nabla \cdot (\boldsymbol{u} + \boldsymbol{v}) &= \nabla \cdot \boldsymbol{u} + \nabla \cdot \boldsymbol{v} && \text{(A.2)} \\ \nabla(u\boldsymbol{v}) &= u\nabla\boldsymbol{v} + \boldsymbol{v}\nabla u && \text{(A.3)} \\ \nabla \cdot u\boldsymbol{v} &= u\,\nabla \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \nabla u && \text{(A.4)} \\ \nabla \times u\boldsymbol{v} &= u\,\nabla \times \boldsymbol{v} + \nabla u \times \boldsymbol{v}\ . && \text{(A.5)} \end{aligned}$$

Note that $\nabla(u\boldsymbol{v})$ is a three-component vector in which the components are three-vectors (with complex values). This means that you have to interpret $\boldsymbol{v}\nabla u$ on the right-hand side of Equation A.3 as a vector (of vectors) constructed by multiplication of each component in the vector ∇u by the vector $\boldsymbol{v}$.

For the insertion in the second order wave equation, we need the the Laplacian and the curl of $\boldsymbol{E}_c$, and for the Laplacian we need the gradient. Using Equation A.3 with Equation 5.5, we have

$$\nabla \boldsymbol{E}_c = e^{i\,k_0 \mathscr{S}} \nabla \boldsymbol{E}_0 + \boldsymbol{E}_0 \nabla e^{i\,k_0 \mathscr{S}}\ . \tag{A.6}$$

The gradient of the exponential part of $\boldsymbol{E}_c$ is

$$\nabla e^{i\,k_0 \mathscr{S}} = i\,k_0 \nabla \mathscr{S}\, e^{i\,k_0 \mathscr{S}}\ . \tag{A.7}$$

Insertion of Equation A.7 in Equation A.6 gives

$$\nabla \boldsymbol{E} = e^{i\,k_0 \mathscr{S}} (\nabla \boldsymbol{E}_0 + i\,k_0 \boldsymbol{E}_0 \nabla \mathscr{S})\ . \tag{A.8}$$

Now the Laplacian is found using Equations A.4, A.2, and A.7 with Equation A.8:

$$\begin{aligned} \nabla^2 \boldsymbol{E}_c &= \nabla \cdot (\nabla \boldsymbol{E}_c) \\ &= e^{i\,k_0 \mathscr{S}} \nabla \cdot (\nabla \boldsymbol{E}_0 + i\,k_0 \boldsymbol{E}_0 \nabla \mathscr{S}) \\ &\quad + (\nabla \boldsymbol{E}_0 + i\,k_0 \boldsymbol{E}_0 \nabla \mathscr{S}) \cdot \nabla e^{i\,k_0 \mathscr{S}} \\ &= e^{i\,k_0 \mathscr{S}} \big(\nabla^2 \boldsymbol{E}_0 + i\,k_0 \nabla \cdot (\boldsymbol{E}_0 \nabla \mathscr{S}) \\ &\quad + i\,k_0 \nabla \boldsymbol{E}_0 \cdot \nabla \mathscr{S} - k_0^2 \boldsymbol{E}_0 \nabla \mathscr{S} \cdot \nabla \mathscr{S}\big) \\ &= e^{i\,k_0 \mathscr{S}} \big(\nabla^2 \boldsymbol{E}_0 + i\,k_0 \boldsymbol{E}_0 \nabla^2 \mathscr{S} \\ &\quad + i\,k_0\, 2\nabla \boldsymbol{E}_0 \cdot \nabla \mathscr{S} - k_0^2 \boldsymbol{E}_0 (\nabla \mathscr{S})^2\big)\ . \end{aligned}$$

The curl is obtained using Equations A.5 and A.7:

$$\begin{aligned} \nabla \times \boldsymbol{E} &= e^{i\,k_0 \mathscr{S}} \nabla \times \boldsymbol{E}_0 + \nabla e^{i\,k_0 \mathscr{S}} \times \boldsymbol{E}_0 \\ &= e^{i\,k_0 \mathscr{S}} (\nabla \times \boldsymbol{E}_0 + i\,k_0 \nabla \mathscr{S} \times \boldsymbol{E}_0)\ . \end{aligned}$$

We will also need $\nabla(\boldsymbol{E} \cdot \nabla \ln \varepsilon)$. It is obtained using Equation 5.5 (without the time exponential) and Equations A.3 and A.7:

$$\begin{aligned} \nabla(\boldsymbol{E} \cdot \nabla \ln \varepsilon) &= e^{i\,k_0 \mathscr{S}} \nabla(\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon) + (\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon) \nabla e^{i\,k_0 \mathscr{S}} \\ &= e^{i\,k_0 \mathscr{S}} \big(\nabla(\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon) + i\,k_0 \nabla \mathscr{S} (\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon)\big)\ . \end{aligned}$$

In addition we need the cross product between $\nabla \ln \mu$ and $\nabla \times \boldsymbol{E}$. Using the expression for the curl of $\boldsymbol{E}$ and the vector triple product, we have

$$\begin{aligned}\nabla \ln \mu \times (\nabla \times \boldsymbol{E}) &= \nabla \ln \mu \times \left(e^{i\,k_0 \mathscr{S}} (\nabla \times \boldsymbol{E}_0 \right. \\ &\quad \left. + i\,k_0 \nabla \mathscr{S} \times \boldsymbol{E}_0)\right) \\ &= e^{i\,k_0 \mathscr{S}} \big(\nabla \ln \mu \times (\nabla \times \boldsymbol{E}_0) \\ &\quad + i\,k_0 \nabla \mathscr{S} (\nabla \ln \mu \cdot \boldsymbol{E}_0) \\ &\quad - i\,k_0 \boldsymbol{E}_0 (\nabla \ln \mu \cdot \nabla \mathscr{S})\big) \ .\end{aligned}$$

Each term of the second order wave equation (5.1) has now been expanded such that insertion of the wave function (5.5) is straightforward. After a few rearrangements, and application of the distributive property of the gradient as well as the dot product, the result is

$$\begin{aligned}& e^{i\,k_0 \mathscr{S}} \big[(i\,k_0)^2 \left((\nabla \mathscr{S})^2 - n^2\right) \boldsymbol{E}_0 \\ &\quad + i\,k_0 \big((\nabla^2 \mathscr{S} - \nabla \ln \mu \cdot \nabla \mathscr{S}) \boldsymbol{E}_0 \\ &\quad + (\boldsymbol{E}_0 \cdot \nabla \ln(\mu\,\varepsilon))\, \nabla \mathscr{S} + 2 \nabla \boldsymbol{E}_0 \cdot \nabla \mathscr{S}\big) \\ &\quad + \nabla^2 \boldsymbol{E}_0 + \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_0) + \nabla (\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon)\big] = \boldsymbol{0} \ .\end{aligned}$$

By division of the equation with $e^{i\,k_0 \mathscr{S}} (i\,k_0)^2$, we obtain

$$\left((\nabla \mathscr{S})^2 - n^2\right) \boldsymbol{E}_0 + (i\,k_0)^{-1} \boldsymbol{L}(\boldsymbol{E}_0, \mathscr{S}, \varepsilon, \mu) + (i\,k_0)^{-2} \boldsymbol{M}(\boldsymbol{E}_0, \varepsilon, \mu) = \boldsymbol{0} \ ,$$

where

$$\begin{aligned}\boldsymbol{L}(\boldsymbol{E}_0, \mathscr{S}, \varepsilon, \mu) &= (\nabla^2 \mathscr{S} - \nabla \ln \mu \cdot \nabla \mathscr{S}) \boldsymbol{E}_0 \\ &\quad + (\boldsymbol{E}_0 \cdot \nabla \ln(\mu\,\varepsilon))\, \nabla \mathscr{S} + 2 \nabla \boldsymbol{E}_0 \cdot \nabla \mathscr{S} \\ \boldsymbol{M}(\boldsymbol{E}_0, \varepsilon, \mu) &= \nabla^2 \boldsymbol{E}_0 + \nabla \ln \mu \times (\nabla \times \boldsymbol{E}_0) \\ &\quad + \nabla (\boldsymbol{E}_0 \cdot \nabla \ln \varepsilon) \ .\end{aligned}$$

The fundamental assumption in geometrical optics is that λ is very small. But then $k_0 = 2\pi/\lambda$ is very large, why the $\boldsymbol{L}$ and $\boldsymbol{M}$ terms are negligible. The resulting equation is then

$$\left((\nabla \mathscr{S})^2 - n^2\right) \boldsymbol{E}_0 = \boldsymbol{0}$$

which reduces to the eikonal equation:

$$(\nabla \mathscr{S})^2 = n^2 \ .$$

Thus we have shown that the eikonal equation is valid for geometrical optics in an inhomogeneous medium with a complex index of refraction.

A.3 The Time Average of Poynting's Vector

To find the time average of Poynting's vector $\boldsymbol{S}_{\text{avg}}$ over one period of oscillation $T = 2\pi/\omega$, we need the cross product of $\text{Re}(\boldsymbol{E}_c)$ and $\text{Re}(\boldsymbol{H}_c)$. Inserting from Equations 5.10 and 5.11 and using the distributive property of the cross product, we find

$$\begin{aligned}\text{Re}(\boldsymbol{E}_c) \times \text{Re}(\boldsymbol{H}_c) &= e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}}\big((\boldsymbol{E}_0' \cos\theta) \times (\boldsymbol{H}_0' \cos\theta - \boldsymbol{H}_0'' \sin\theta) \\ &\quad - (\boldsymbol{E}'' \sin\theta) \times (\boldsymbol{H}_0' \cos\theta - \boldsymbol{H}_0'' \sin\theta)\big) \\ &= e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}}\big((\boldsymbol{E}_0' \times \boldsymbol{H}_0') \cos^2\theta - (\boldsymbol{E}_0' \times \boldsymbol{H}_0'' \\ &\quad + \boldsymbol{E}_0'' \times \boldsymbol{H}_0') \cos\theta \sin\theta + (\boldsymbol{E}_0'' \times \boldsymbol{H}_0'') \sin^2\theta\big) \ .\end{aligned}$$

Integration of $\text{Re}(\boldsymbol{E}_c) \times \text{Re}(\boldsymbol{H}_c)$ over the period of oscillation $T = 2\pi/\omega$ gives

$$\begin{aligned}\int_0^T \text{Re}(\boldsymbol{E}_c) \times \text{Re}(\boldsymbol{H}_c)\,\mathrm{d}t &= e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}}\Big[(\boldsymbol{E}_0' \times \boldsymbol{H}_0')\left(\frac{t}{2} - \frac{\cos\theta\sin\theta}{2\omega}\right) \\ &\quad - (\boldsymbol{E}_0' \times \boldsymbol{H}_0'' + \boldsymbol{E}_0'' \times \boldsymbol{H}_0') \cos^2\theta \\ &\quad + (\boldsymbol{E}_0'' \times \boldsymbol{H}_0'')\left(\frac{t}{2} + \frac{\cos\theta\sin\theta}{2\omega}\right)\Big]_0^T \ .\end{aligned}$$

Since the integral of a sine or a cosine function over its period is zero (and $\theta = \omega t - \boldsymbol{k}' \cdot \boldsymbol{x}$), the time average of Poynting's vector now follows:

$$\begin{aligned}\boldsymbol{S}_{\text{avg}} &= \frac{\varepsilon_0 c^2 \mu}{T} \int_0^T \text{Re}(\boldsymbol{E}_c) \times \text{Re}(\boldsymbol{H}_c)\,\mathrm{d}t \\ &= \frac{\varepsilon_0 c^2 \mu}{2} (\boldsymbol{E}_0' \times \boldsymbol{H}_0' + \boldsymbol{E}_0'' \times \boldsymbol{H}_0'')\, e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ . \qquad \text{(A.9)}\end{aligned}$$

A similar result was found by Bell [1967, Eqn. 7.2], but he gave no details about the derivation.

To find the direction of Poynting's vector it is necessary that we analyze $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$. Considering the definition of the index of refraction n (4.41) and the wave number in vacuum $k_0 = \omega/c$, another way to write the first two time-harmonic Maxwell equations (4.30–4.31) is

$$\nabla \times \boldsymbol{H}_c = -ik_0(c\mu)^{-1} n^2 \boldsymbol{E}_c \qquad \text{(A.10)}$$

$$\nabla \times \boldsymbol{E}_c = ik_0 c\mu \boldsymbol{H}_c \ . \qquad \text{(A.11)}$$

Inserting the wave functions (5.5-5.6), dividing out the time exponential, and using Equation A.5 to expand the curls, we obtain

$$\begin{aligned}e^{ik_0\mathscr{S}}(\nabla \times \boldsymbol{H}_0 + ik_0 \nabla\mathscr{S} \times \boldsymbol{H}_0) &= -ik_0(c\mu)^{-1} n^2 \boldsymbol{E}_0 e^{ik_0\mathscr{S}} \\ e^{ik_0\mathscr{S}}(\nabla \times \boldsymbol{E}_0 + ik_0 \nabla\mathscr{S} \times \boldsymbol{E}_0) &= ik_0 c\mu \boldsymbol{H}_0 e^{ik_0\mathscr{S}} \ .\end{aligned}$$

Division by $ik_0 e^{ik_0\mathscr{S}}$ at both sides of both equations gives

$$\begin{aligned}\frac{1}{i\,k_0}\nabla\times\boldsymbol{H}_0+\nabla\mathscr{S}\times\boldsymbol{H}_0 &= -(c\mu)^{-1}n^2\boldsymbol{E}_0\\ \frac{1}{i\,k_0}\nabla\times\boldsymbol{E}_0+\nabla\mathscr{S}\times\boldsymbol{E}_0 &= c\mu\boldsymbol{H}_0\ .\end{aligned}$$

The fundamental assumption of geometrical optics ($\lambda\to 0$), means that $k_0 = 2\pi/\lambda$ is very large, and for that reason the first term is negligible in both equations. Hence,

$$\boldsymbol{E}_0 = -c\mu\, n^{-2}\nabla\mathscr{S}\times\boldsymbol{H}_0 \tag{A.12}$$

$$\boldsymbol{H}_0 = (c\mu)^{-1}\nabla\mathscr{S}\times\boldsymbol{E}_0\ . \tag{A.13}$$

For a transverse electric (TE) wave, we have $\boldsymbol{E}_0 = \boldsymbol{E}_0'$. Then Equation A.9 reduces to

$$\boldsymbol{S}_{\mathrm{avg,TE}} = \frac{\varepsilon_0 c^2\mu}{2}\,(\boldsymbol{E}_0\times\boldsymbol{H}_0')\,e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}}\ .$$

To find $\boldsymbol{E}_0\times\boldsymbol{H}_0'$, we insert the real part of $\boldsymbol{H}_0$ (A.13) and use the fact that $\boldsymbol{E}_0$ is real as well as the vector triple product:

$$\begin{aligned}\boldsymbol{E}_0\times\boldsymbol{H}_0' &= \mathrm{Re}\left((c\mu)^{-1}\boldsymbol{E}_0\times(\nabla\mathscr{S}\times\boldsymbol{E}_0)\right)\\ &= \mathrm{Re}\left((c\mu)^{-1}\left((\boldsymbol{E}_0\cdot\boldsymbol{E}_0)\nabla\mathscr{S}-(\boldsymbol{E}_0\cdot\nabla\mathscr{S})\boldsymbol{E}_0\right)\right)\ .\end{aligned}$$

To analyze $\boldsymbol{E}_0\cdot\nabla\mathscr{S}$ we return to Maxwell's equations. Inserting the wave function (5.5) in the third equation (4.32) and expanding it using Equation A.4 gives

$$\nabla\cdot(\varepsilon\boldsymbol{E}_c) = e^{-i\omega t}\left(\varepsilon e^{ik_0\mathscr{S}}\nabla\cdot\boldsymbol{E}_0+\boldsymbol{E}_0\cdot\nabla(\varepsilon e^{ik_0\mathscr{S}})\right)=0\ .$$

Then using a variant of Equation A.3 as well as Equation A.7, we get

$$e^{ik_0\mathscr{S}}\left(\varepsilon\nabla\cdot\boldsymbol{E}_0+\boldsymbol{E}_0\cdot\nabla\varepsilon+ik_0\varepsilon\boldsymbol{E}_0\cdot\nabla\mathscr{S}\right)=0\ .$$

Division by $ik_0\varepsilon e^{i\,k_0\mathscr{S}}$ at both sides and using $u^{-1}\nabla u = \nabla\ln u$ gives

$$\frac{1}{ik_0}\left(\nabla\cdot\boldsymbol{E}_0+\boldsymbol{E}_0\cdot\nabla\ln\varepsilon\right)+\boldsymbol{E}_0\cdot\nabla\mathscr{S}=0\ .$$

Now, since k_0 is assumed very large in geometrical optics, the first term is neglected revealing

$$\boldsymbol{E}_0\cdot\nabla\mathscr{S}=0\ . \tag{A.14}$$

Using the same line of arguments, but starting from the fourth Maxwell equation (4.33), it also follows that

$$\boldsymbol{H}_0\cdot\nabla\mathscr{S}=0\ . \tag{A.15}$$

In light of Equation A.14 we see that

$$\boldsymbol{E}_0 \times \boldsymbol{H}_0' = \mathrm{Re}\left((c\mu)^{-1}(\boldsymbol{E}_0 \cdot \boldsymbol{E}_0)\nabla\mathscr{S}\right) = (c\mu)^{-1}|\boldsymbol{E}_0|^2\mathrm{Re}(\nabla\mathscr{S}) \ ,$$

but then

$$\boldsymbol{S}_{\mathrm{avg,TE}} = \frac{\varepsilon_0 c}{2}|\boldsymbol{E}_0|^2\mathrm{Re}(\nabla\mathscr{S})e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ ,$$

which shows that a ray representing a TE wave follows the direction given by the real part of $\nabla\mathscr{S}$.

For a transverse magnetic (TM) wave we have $\boldsymbol{H}_0 = \boldsymbol{H}_0'$. Then Equation A.9 reduces to

$$\boldsymbol{S}_{\mathrm{avg,TM}} = \frac{\varepsilon_0 c^2 \mu}{2}\left(\boldsymbol{E}_0' \times \boldsymbol{H}_0\right) e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ .$$

This time we we insert the real part of $\boldsymbol{E}_0$ (A.12) and use the fact that $\boldsymbol{H}_0$ is real as well as the vector triple product:

$$\begin{aligned} \boldsymbol{E}_0' \times \boldsymbol{H}_0 &= \mathrm{Re}\left(-c\mu n^{-2}(\nabla\mathscr{S} \times \boldsymbol{H}_0) \times \boldsymbol{H}_0\right) \\ &= \mathrm{Re}\left(-c\mu n^{-2}\left((\nabla\mathscr{S} \cdot \boldsymbol{H}_0)\boldsymbol{H}_0 - (\boldsymbol{H}_0 \cdot \boldsymbol{H}_0)\nabla\mathscr{S}\right)\right) . \end{aligned}$$

According to Equation A.15 the first term vanishes and we get

$$\boldsymbol{S}_{\mathrm{avg,TM}} = \frac{\varepsilon_0 c^3 \mu^2}{2}|\boldsymbol{H}_0|^2\mathrm{Re}\left(\frac{\nabla\mathscr{S}}{n^2}\right) e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ .$$

Keeping in mind that the index of refraction $n = n' + in''$ is a complex number, we observe that

$$\begin{aligned} n^2 &= n'^2 - n''^2 + i\,2n'n'' \\ |n^2|^2 &= \left(n'^2 - n''^2\right)^2 + 4n'^2n''^2 = \left(n'^2 + n''^2\right)^2 \ . \end{aligned}$$

Another way to write the TM component of Poynting's vector is therefore

$$\boldsymbol{S}_{\mathrm{avg,TM}} = \frac{\varepsilon_0 c^3 \mu^2}{2}|\boldsymbol{H}_0|^2\frac{(n'^2 - n''^2)\mathrm{Re}(\nabla\mathscr{S}) + 2n'n''\mathrm{Im}(\nabla\mathscr{S})}{(n'^2 + n''^2)^2}e^{-2\boldsymbol{k}''\cdot\boldsymbol{x}} \ .$$

This means that a ray representing a TM wave follows the direction given by the combination of the index of refraction and the real and imaginary parts of $\nabla\mathscr{S}$ found above.

Similar results for the direction of the time averaged Poynting vector for TE and TM waves in homogeneous media were found by Bell [1967], but no details were given about the derivation. The derivation given here is also valid for inhomogeneous media.

Appendix B

On Polynomial Arrays

B.1 raisedegree

To define **raisedegree**, we need the following standard array-theoretic operator:

- EACHBOTH.

And we need the following standard array-theoretic operations:

- **pass**
- **first**
- **last**
- **front**
- **rest**
- **tally**
- **reverse**
- **append**

- **hitch**
- **grid**
- **div**.

Definitions of these operators and operations are available in several references on array theory [More 1981; Pedersen and Hansen 1988; Falster 1997; Jenkins and Falster 1999; Nial 2006].

The definition of **raisedegree** follows the formulae [Gravesen 2002]:

$$\hat{\boldsymbol{b}}_0 = \boldsymbol{b}_0 \tag{B.1}$$

$$\hat{\boldsymbol{b}}_k = \frac{n+1-k}{n+1}\boldsymbol{b}_k + \frac{k}{n+1}\boldsymbol{b}_{k-1} \quad , \quad k = 1, \ldots, n \tag{B.2}$$

$$\hat{\boldsymbol{b}}_{n+1} = \boldsymbol{b}_n \ , \tag{B.3}$$

where n is the degree of the Bézier curve. First, we define operations to find the two terms in Equation B.2:

$$\begin{aligned} \text{elevweight1} &= \text{EACHBOTH}(\cdot)\ [\text{div}\ [1 + \text{reverse grid}, \text{tally}], \text{pass}] \\ \text{elevweight2} &= \text{EACHBOTH}(\cdot)\ [\text{div}\ [1 + \text{grid}, \text{tally}], \text{pass}] \ . \end{aligned}$$

Then we add the terms to find the middle part of the elevated array

$$\text{elevmiddle} = \text{EACHBOTH}(+)\ [\text{rest elevweight1}, \text{front elevweight2}] \ .$$

Finally, the operation **raisedegree** is obtained by linking the middle to the end control points (Equations B.1–B.3). The definition is

$$\text{raisedegree} = \text{append}\ [\text{hitch}\ [\text{first}, \text{elevmiddle}], \text{last}] \ .$$

B.2 polyplace

To define **polyplace**, we need the following standard array-theoretic operators:

- EACHLEFT
- EACHRIGHT
- CONVERSE.

And we need the following standard array-theoretic operations:

- **pass**
- **second**
- **last**
- **sublist**
- **reshape**
- **grid**.

Definitions of these operators and operations are available in several references on array theory [More 1981; Pedersen and Hansen 1988; Falster 1997; Jenkins and Falster 1999; Nial 2006].

Let us define a couple of operations to help the definition of **polyplace**. First we define an operation which constructs an array holding the grid of the colligated array:

$$\text{gridlast} = \text{grid}\ (0\ \text{CONVERSE}(\text{reshape}))\ (1 + \text{last})\ .$$

Then we find the items to be placed in the colligated array using **polyindex** (which was defined in Section 14.2). The operation to do this is defined by

$$\text{findgridequals} = \text{EACHLEFT}(\text{EACHRIGHT}(=))\ [\text{gridlast}, \text{pass}]\ \text{polyindex}\ .$$

Finally, the operation **polyplace** is defined by

$$\text{polyplace} = \text{EACHLEFT}(\text{sublist})\ [\text{findgridequals}, \text{second}]\ .$$

Appendix C

On the Milk Model

For some applications, it is convenient to have a functional expression for the macroscopic optical properties of a material. Using the computations described in Part II, it is not difficult to sample the macroscopic optical properties for milk with many different fat and protein contents. We may even sample RGB values using the colour matching functions discussed in Chapter 11. To get a set of functional expressions for the macroscopic optical properties, we do a number of least squares fits of the samples. This gives the following set of RGB vector functions mapping fat and protein content of milk directly to its macroscopic optical properties:

$$\sigma_a = \begin{bmatrix} 1.381 - 0.008600 w_p + 1.209 w_f \\ 2.201 - 0.01459 w_p + 1.982 w_f \\ 10.13 - 0.07048 w_p + 4.170 w_f \end{bmatrix} \tag{C.1}$$

$$\sigma_s = \begin{bmatrix} 213.5 w_p + 15631 w_f^{1.24} e^{h_1(\ln(w_f))_R} \\ 338.3 w_p + 18349 w_f^{1.15} e^{h_1(\ln(w_f))_G} \\ 614.0 w_p + 22585 w_f^{1.01} e^{h_1(\ln(w_f))_B} \end{bmatrix} \tag{C.2}$$

$$g = \begin{bmatrix} (18.63 w_p + (\sigma_{sR} - 213.5 w_p)\tilde{g}(w_f)_R)/\sigma_{sR} \\ (37.79 w_p + (\sigma_{sG} - 338.3 w_p)\tilde{g}(w_f)_G)/\sigma_{sG} \\ (96.69 w_p + (\sigma_{sB} - 614.0 w_p)\tilde{g}(w_f)_B)/\sigma_{sB} \end{bmatrix}. \tag{C.3}$$

A two-piece fit is needed for the asymmetry parameter, and consequently the $\tilde{g}$ function in Equation C.3 is given by

$$\tilde{g}(0 < w_f < 0.7) = \begin{bmatrix} 0.9523 w_f^{-0.0120} e^{h_2(\ln(1/w_f))_R} \\ 0.9539 w_f^{-0.00783} e^{h_2(\ln(1/w_f))_G} \\ 0.9554 w_f^{-0.000161} e^{h_2(\ln(1/w_f))_B} \end{bmatrix} \tag{C.4}$$

$$\tilde{g}(w_f \geq 0.7) = \begin{bmatrix} 0.9576 w_f^{0.00911} e^{h_3(\ln(w_f))_R} \\ 0.9585 w_f^{0.00783} e^{h_3(\ln(w_f))_G} \\ 0.9577 w_f^{0.00531} e^{h_3(\ln(w_f))_B} \end{bmatrix} . \tag{C.5}$$

If $w_f = 0$, replace $\tilde{g}$ by $\mathbf{0}$ and $\ln(w_f)$ by 0. Finally the functions h_1, h_2, and h_3 are the following RGB vector polynomials

$$h_1(x) = \begin{bmatrix} -0.00129x^4 + 0.0305x^3 - 0.219x^2 \\ -0.00149x^4 + 0.0327x^3 - 0.213x^2 \\ -0.00206x^4 + 0.0373x^3 - 0.202x^2 \end{bmatrix} \tag{C.6}$$

$$h_2(x) = \begin{bmatrix} -0.0386x^3 - 0.00543x^2 \\ -0.0368x^3 + 0.00266x^2 \\ -0.0334x^3 + 0.0111x^2 \end{bmatrix} \tag{C.7}$$

$$h_3(x) = \begin{bmatrix} 0.000281x^3 - 0.00366x^2 \\ 0.000379x^3 - 0.00401x^2 \\ 0.000509x^3 - 0.00429x^2 \end{bmatrix} . \tag{C.8}$$

The absorption and scattering coefficients in this appearance model have the unit m^{-1}. For every band of the three optical properties (C.1,C.2,C.3) the maximum error is 10.2%. This error band only excludes input parameters where $w_f < 0.05$ wt.-%. The maximum error occurs in the blue band of the absorption coefficient. For the majority of the possible input parameters, the error rarely exceeds 2% in all bands of all properties. The fit is best in the region where we normally find the fat and protein contents of milk.

Bibliography

Al-Kindī, ∼870, 1997. De aspectibus. In R. Rashed *Œuvres Philosophiques et Scientifiques D'Al-Kindi: L'Optique & la Catoprique*, E. J. Brill, 1997. English excerpt by Smith [1999].

Amanatides, J. 1984. Ray tracing with cones. *Computer Graphics (Proceedings of ACM SIGGRAPH 84) 18*, 3 (July), 129–135.

Antonov, J. I., Locarnini, R. A., Boyer, T. P., Mishonov, A. V., and Garcia, H. E. 2006. Salinity. In *NOAA Atlas NESDIS 62*, S. Levitus, Ed., vol. 2 of *World Ocean Atlas 2005*. U.S. Government Printing Office, Washington, D.C. CD-ROM.

Appel, A. 1968. Some techniques for shading machine renderings of solids. In *AFIPS 1968 Spring Joint Computer Conference Proceedings*, vol. 32, 37–45.

Aristotle, ∼350 B.C., 1941. De Anima (On the Soul). Translated by J. A. Smith in *The Basic Works of Aristotle*, Richard McKeon ed., Random House, 1941.

Aristotle, ∼350 B.C., 1984. Meteorology. Translated by E. W. Webster in *The Complete Works of Aristotle*, J. Barnes ed., Princeton University Press, 1984.

Aronson, R., and Corngold, N. 1999. Photon diffusion coefficient in an absorbing medium. *Journal of the Optical Society of America A 16*, 5 (May), 1066–1071.

Arvo, J. 1986. Backward ray tracing. In *Developments in Ray Tracing*, ACM SIGGRAPH 86 Course Notes, ACM Press.

Askebjer, P., Barwick, S. W., Bergström, L., Bouchta, A., Carius, S., Dalberg, E., Engel, K., Erlandsson, B., Goobar, A., Gray, L.,

Hallgren, A., Halzen, F., Heukenkamp, H., Hulth, P. O., Hundertmark, S., Jacobsen, J., Karle, A., Kandhadai, V., Liubarsky, I., Lowder, D., Miller, T., Mock, P., Morse, R. M., Porrata, R., Buford Price, P., Richards, A., Rubinstein, H., Schneider, E., Spiering, C., Streicher, O., Sun, Q., Thon, T., Tilav, S., Wischnewski, R., Walck, C., and Yodh, G. B. 1997. Optical properties of deep ice at the South Pole: Absorption. *Applied Optics 36*, 18 (June), 4168–4180.

Assur, A. 1958. Composition of sea ice and its strength. In *Arctic Sea Ice*. U. S. National Academy of Sciences, Washington, D. C., 106–138. National Research Council Publication 598.

Attaie, R., and Richtert, R. L. 2000. Size distribution of fat globules in goat milk. *Journal of Dairy Science 83*, 940–944.

Babin, M., Morel, A., Fell, V. F.-S. F., and Stramski, D. 2003. Light scattering properties of marine particles in coastal and open ocean waters as related to the particle mass concentration. *Limnology and Oceanography 48*, 2, 843–859.

Babin, M., Stramski, D., Ferrari, G. M., Claustre, H., Bricaud, A., Obolensky, G., and Hoepffner, N. 2003. Variations in the light absorption coefficients of phytoplankton, nonalgal particles, and dissolved organic matter in coastal waters around Europe. *Journal of Geophysical Research 108*, C7, 3211 (July), **4**–1–20.

Bahar, E., and Chakrabarti, S. 1987. Full-wave theory applied to computer-aided graphics for 3D objects. *IEEE Computer Graphics & Applications 7*, 7 (July), 46–60.

Banks, D. C., and Abu-Raddad, L. 2007. The foundations of photo-realistic rendering: From quantum electrodynamics to Maxwell's equations. In *Proceedings of the IASTED International Conference on Graphics and Visualization in Engineering*, M. S. Alam, Ed.

Bartholin, R. 1670. *Experimenta crystalli Islandici disdiaclastici: Quibus mira & insolita refractio detegitur*. Danielis Paulli.

Becquerel, E. 1868. *La Lumière, ses causes et ses effects*. Firmin Didot Frères. Tome Second: Effects de la lumière.

Bell, E. E. 1967. Optical constants and their measurement. In *Light and Matter Ia*, S. Flügge and L. Genzel, Eds., vol. XXV/2a of *Handbuch der Physik (Encyclopedia of Physics)*. Springer-Verlag, Berlin, ch. 1, 1–58.

Belokopytov, G. V., and Vasil'ev, E. N. 2006. Scattering of a plane inhomogeneous electromagnetic wave by a spherical particle. *Radiophysics*

and Quantum Electronics 49, 1 (January), 65–73. Translated from Izvestiya Vysshikh Uchebnykh Zavedenii, Radiofizika, Vol. 49, No. 1, pp. 72–81, January 2006.

Bennett, J. A. 1974. Complex rays for radio waves in an absorbing ionosphere. *Proceedings of the IEEE 62*, 11 (November), 1577–1585.

Berger, M., and Trout, T. 1990. Ray tracing mirages. *IEEE Computer Graphics & Applications 10*, 3 (May), 36–41.

Blinn, J. F. 1977. Models of light reflections for computer synthesized pictures. *Proceedings of ACM SIGGRAPH 77* (July), 192–198.

Blinn, J. F. 1982. Light reflection functions for simulation of clouds and dusty surfaces. *Computer Graphics (Proceedings of ACM SIGGRAPH 82) 16*, 3 (July), 21–29.

Bohr, N. 1913. On the constitution of atoms and molecules. *Philosophical Magazine 26*, 1–25.

Bohren, C. F., and Gilra, D. P. 1979. Extinction by a spherical particle in an absorbing medium. *Journal of Colloid and Interface Science 72*, 2 (November), 215–221.

Bohren, C. F., and Huffman, D. R. 1983. *Absorption and Scattering of Light by Small Particles*. John Wiley & Sons, Inc.

Bohren, C. F. 1983. Colors of snow, frozen waterfalls, and icebergs. *Journal of Optical Society America 73*, 12 (December), 1646–1652.

Boltzmann, L. 1884. Ableitung des Stefan'schen Gesetzes, betreffend die Abhängigkeit der Wärmestrahlung von der Temperatur aus der elektromagnetischen Lichttheorie. *Annalen der Physik und Chemie 258*, 6, 291–294.

Born, M., and Wolf, E. 1999. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, seventh (expanded) ed. Cambridge University Press.

Bothe, W. 1941. Einige Diffusionsprobleme. *Zeitschrift für Physik 118*, 7–8 (July), 401–408.

Bothe, W. 1942. Die Diffusion von einer Punktquelle aus (Nachtrag der Arbeit "Einige Diffusionsprobleme"). *Zeitschrift für Physik 119*, 7–8 (July), 493–497.

Bouguer, P., 1729. Essai d'optique sur la gradation de la lumiere. Reprinted in *Les maîtres de la pensée scientifique*, Gauthier-Villars, 1921. An extended version was published posthumously as *Traité d'Optique sur la gradation de la lumiere: Ouvrage posthume de M. Bouguer, de l'Académie Royale des Sciences, &c.*, H. L. Guerin & L. F. Delatour, 1760.

Bresenham, J. E. 1965. Algorithm for computer control of a digital plotter. *IBM Systems Journal 4*, 1, 25–30.

Bricaud, A., Babin, M., Morel, A., and Claustre, H. 1995. Variability in the chlorophyll-specific absorption coefficients of natural phytoplankton: Analysis and parameterization. *Journal of Geophysical Research 100*, C7 (July), 13321–13332.

Cachorro, V. E., and Salcedo, L. L. 1991. New improvements for Mie scattering calculations. *Journal of Electromagnetic Waves and Applications 5*, 9, 913–926.

Callet, P. 1996. Pertinent data for modelling pigmented materials in realistic rendering. *Computer Graphics Forum 15*, 2, 119–127.

Case, K. M., and Zweifel, P. F. 1967. *Linear Transport Theory*. Addison-Wesley Publishing Company.

Champeney, D. C. 1973. *Fourier Transforms and their Physical Applications*. Academic Press, London.

Chandrasekhar, S. 1950. *Radiative Transfer*. Oxford, Clarendon Press. Unabridged and slightly revised version published by Dover Publications, Inc., in 1960.

Chapman, S. J., Lawry, J. M. H., Ockendon, J. R., and Tew, R. H. 1999. On the theory of complex rays. *SIAM Review 41*, 3 (September), 417–509.

Clausius, R. 1879. *Mechanische Wärmetheorie*, 2nd ed., vol. 2. Braunschweig.

Cohen, E., Riesenfeld, R. F., and Elber, G. 2001. *Geometric Modeling with Splines: An Introduction*. A K Peters, Natick, Massachusetts.

Compton, A. H. 1923. A quantum theory of the scattering of X-rays by light elements. *Physical Review 21*, 5 (May), 483–502.

Cook, R. L., Porter, T., and Carpenter, L. 1984. Distributed ray tracing. *Computer Graphics (Proceedings of ACM SIGGRAPH 84) 18*, 3 (July), 137–145.

Cox, G. F. N., and Weeks, W. F. 1983. Equations for determining the gas and brine volumes in sea-ice samples. *Journal of Glaciology 29*, 102, 306–316.

Dave, J. V. 1969. Scattering of electromagnetic radiation by a large, absorbing sphere. *IBM Journal of Research and Development 13*, 3 (May), 302–313.

Debye, P. 1909. Der Lichtdruck auf Kugeln von beliebigem Material. *Annalen der Physik 335*, 11, 57–136.

Descartes, R. 1637, 2001. *Discourse on Method, Optics, Geometry, and Meteorology*, revised ed. Hackett Publishing Company, Inc. Translated by P. J. Olscamp from Descartes' *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences*, 1637.

Dias, M. L. 1991. Ray tracing interference color. *IEEE Computer Graphics & Applications 11*, 2 (March), 54–60.

Dieckmann, G., Hemleben, C., and Spindler, M. 1987. Biogenic and mineral inclusions in a green iceberg from the weddell sea, antarctica. *Polar Biology 7*, 1, 31–33.

Diocles, ∼190 B.C., 1975. On burning mirrors. Translated by G. J. Toomer in *On Burning Mirrors: The Arabic Translation of the Lost Greek Original*, Springer-Verlag, 1975.

Dirac, P. A. M. 1927. The quantum theory of the emission and absorption of radiation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 114*, 767 (March), 243–265.

Dirac, P. A. M. 1930. *The Principles of Quantum Mechanics*. Oxford University Press. Fourth edition published 1958.

Dirac, P. A. M. 1966. *Lectures on Quantum Field Theory*. Belfer Graduate School of Science, Yeshiva University, New York.

Dorsey, J., and Hanrahan, P. 1996. Modeling and rendering of metallic patinas. In *Proceedings of ACM SIGGRAPH 1996*, ACM Press, 387–396.

Dorsey, J., Pedersen, H. K., and Hanrahan, P. 1996. Flow and changes in appearance. In *Proceedings of ACM SIGGRAPH 1996*, ACM Press, 411–420.

Dorsey, J., Edelman, A., Jensen, H. W., Legakis, J., and Pedersen, H. K. 1999. Modeling and rendering of weathered stone. In *Proceedings of ACM SIGGRAPH 1999*, ACM Press, 411–420.

Du, H., Fuh, R.-C. A., Li, J., Corkan, L. A., and Lindsey, J. S. 1998. PhotochemCAD: A computer-aided design and research tool in photochemistry. *Photochemistry and Photobiology 68*, 2, 141–142.

Egan, W. G., and Hilgeman, T. W. 1979. *Optical Properties of Inhomogeneous Materials*. Academic Press, New York.

Einstein, A. 1905. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik 322*, 132–148.

Einstein, A. 1906. Zur Theorie der Lichterzeugung und Lichtabsorption. *Annalen der Physik 325*, 199–206.

Einstein, A. 1916. Zur Quantentheorie der Strahlung. *Mitteilungen der Physikalischen Gesellschaft Zürich 16*, 47–62.

Elaloufi, R., Carminati, R., and Greffet, J.-J. 2003. Definition of the diffusion coefficient in scattering and absorbing media. *Journal of the Optical Society of America A 20*, 4 (April), 678–685.

Epstein, P. S. 1930. Geometrical optics in absorbing media. *Proceedings of the National Academy of Sciences 16*, 1 (January), 37–45.

Euclid, ∼300 B.C., 1895. Catoptrics. In J. L. Heiberg, ed., *Euclidis opera omnia*, Vol. 7, pp. 285–343, Teubner, 1895. English excerps by Smith [1999].

Euclid, ∼300 B.C., 1945. Optics. Translated by H. E. Burton in *Journal of the Optical Society of America*, 35:5, pp. 357–372, 1945.

Euler, L., 1746. Nova theoria lucis & colorum. In *Oposcula varii argumenti*, Vol. 1, pp. 169–244, A. Haude & J. C. Spener, 1746.

Falster, P. 1997. Array theory and the definition sequence. Tech. rep., Electric Power Engineering Department, Technical University of Denmark, May.

Fante, R. L. 1981. Relationship between radiative-transport theory and Maxwell's equations in dielectric media. *Journal of the Optical Society of America 71*, 4 (April), 460–468.

Farrell, T. J., Patterson, M. S., and Wilson, B. 1992. A diffusion theory model of spatially resolved, steady-state diffuse reflectance for the noninvasive determination of tissue optical properties *in vivo*. *Medical Physics 19*, 4 (July/August), 879–888.

Fermat, P. d. 1891–1912. *Œuvres de Fermat: Publiées par les soins de MM. Paul Tannery et Charles Henry sous les auspices du Ministère de l'instruction publique*, vol. 2. Gauthier-Villars et fils. Correspondance.

Feynman, R. P., Leighton, R. B., and Sands, M. 1963. *The Feynman Lectures on Physics: Mainly Mechanics, Radiation, and Heat*. Addison-Wesley Publishing Company, Reading, Massachusetts. The Definitive Edition published by Pearson Addison Wesley in 2006.

Feynman, R. P., Leighton, R. B., and Sands, M. 1964. *The Feynman Lectures on Physics: Mainly Electromagnetism and Matter*. Addison-Wesley Publishing Company, Reading, Massachusetts. The Definitive Edition published by Pearson Addison Wesley in 2006.

Feynman, R. P., Leighton, R. B., and Sands, M. 1965. *The Feynman Lectures on Physics: Quantum Mechanics*. Addison-Wesley Publishing Company, Reading, Massachusetts. The Definitive Edition published by Pearson Addison Wesley in 2006.

Feynman, R. P. 1949. The theory of positrons. *Physical Review 76*, 6 (September), 749–759.

Feynman, R. P. 1949. Space-time approach to quantum electrodynamics. *Physical Review 76*, 6 (September), 769–789.

Feynman, R. P. 1985. *QED: The Strange Theory of Light and Matter*. Princeton University Press.

Fick, A. 1855. On liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science X*, 30–39. Abstracted by the author from the German original: Über Diffusion, Poggendorff's Annalen der Physik und Chemie, Vol. 94, pp. 59–86, 1855.

Fox, P. F., and McSweeney, P. L. H. 1998. *Dairy Chemistry and Biochemistry*. Blackie Academic & Professional, London.

Fraenkel, A. A. 1953. *Abstract Set Theory*. North-Holland Publishing Company, Amsterdam, The Netherlands.

Franksen, O. I., and Falster, P. 2000. Colligation or, the logical inference of interconnection. *Mathematics and Computers in Simulation 52*, 1 (March), 1–9.

Franksen, O. I. 1979. Group representation of finite polyvalent logic: A case study using APL notation. In *A Link between Science and Applications of Automatic Control, IFAC VII, World Congress 1978*, Pergamon Press, Oxford, A. Niemi, Ed., 875–887.

Franksen, O. I. 1984. Are data structures geometrical objects? I: Invoking the erlanger program. *Systems Analysis Modelling Simulation 1*, 2, 113–130.

Franksen, O. I. 1984. Are data structures geometrical objects? II: Invariant forms in APL and beyond. *Systems Analysis Modelling Simulation 1*, 2, 131–150.

Franksen, O. I. 1984. Are data structures geometrical objects? III: Appendix A. Linear differential operators. *Systems Analysis Modelling Simulation 1*, 3, 251–260.

Franksen, O. I. 1984. Are data structures geometrical objects? IV: Appendix B: Logic invariants by finite truth-tables. *Systems Analysis Modelling Simulation 1*, 4, 339–350.

Fresnel, A. 1816. Sur la diffraction de la lumière, ou l'on examine particulièrement le phénomène des franges colorées que présentent les ombres des corps éclairés pur un point lumineux. *Annales de Chimie et de Physique 1*, 2, 239–281.

Fresnel, A. 1827. Mémoire sur la double réfraction. *Mémoires de l'Académie des sciences de l'Institut de France 7*, 45–176. Collection of three memoirs presented at 26 November 1821, 22 January 1822, and 26 April 1822.

Fresnel, A. 1832. Mémoire sur la loi des modifications que la réflexion imprime a la lumière polarisée. *Mémoires de l'Académie des sciences de l'Institut de France 11*, 393–434. Presented 7 January 1823.

Frisvad, J. R., Christensen, N. J., and Jensen, H. W. 2007. Computing the scattering properties of participating media using lorenz-mie theory. *ACM Transactions on Graphics 26*, 3 (July). Article 60.

Fu, Q., and Sun, W. 2006. Apparent optical properties of spherical particles in absorbing medium. *Journal of Quantitative Spectroscopy and Radiative Transfer 100*, 1-3, 137–142.

Galen, ∼180 A.D., 1984. On the doctrines of Hippocrates and Plato. In P. H. De Lacy, ed. and trans., *Galeni De placitis Hippocratis et Platonis, Corpus Medicorum Graecorum*, V, 4, 1, three volumes, 1978–1984.

Gebhardt, R., Doster, W., Friedrich, J., and Kulozik, U. 2006. Size distribution of pressure-decomposed casein micelles studied by dynamic light scattering and AFM. *European Biophysics Journal 35*, 503–509.

Glassner, A. S. 1995. *Principles of Digital Image Synthesis*. Morgan Kaufmann, San Francisco, California. Two-volume set.

Glasstone, S., and Edlund, M. C. 1952. *The Elements of Nuclear Reactor Theory*. D. van Nostrand Company, Inc., Princeton, New Jersey.

Goedecke, G. H. 1977. Radiative transfer in closely packed media. *Journal of the Optical Society of America 67*, 10 (October), 1339–1348.

Goff, H. D., and Hill, A. R. 1993. Chemistry and physics. In *Dairy Science and Technology Handbook: Principles and Properties*, Y. H. Hui, Ed., vol. 1. VCH Publishers, Inc., New York, ch. 1, 1–81.

Goral, C. M., Torrance, K. E., Greenberg, D. P., and Battaile, B. 1984. Modeling the interaction of light between diffuse surfaces. *Computer Graphics (Proceedings of ACM SIGGRAPH 84) 18*, 3 (July), 213–222.

Gouraud, H. 1971. Continuous shading of curved surfaces. *IEEE Transactions on Computers C-20*, 6 (June).

Gravesen, J. 2002. *Differential Geometry and Design of Shape and Motion*. Department of Mathematics, Technical University of Denmark, November. Lecture notes for Course 01243.

Gray, D. E., Ed. 1972. *American Institute of Physics Handbook*, 3rd ed. McGraw-Hill.

Greene, N. 1986. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications 6*, 11 (November), 21–29.

Grenfell, T. C., and Perovich, D. K. 1981. Radiation absorption coefficients of polycrystalline ice from 400-1400 nm. *Journal of Geophysical Research 86*, C8 (August), 7447–7450.

Grenfell, T. C., and Warren, S. G. 1999. Representation of a nonspherical ice particle by a collection of independent spheres for scattering and absorption of radiation. *Journal of Geophysical Research 104*, D24 (December), 31,697–31,709.

Grenfell, T. C., Neshyba, S. P., and Warren, S. G. 2005. Representation of a non-spherical ice particle by a collection of independent spheres for scattering and absorption of radiation: 3. Hollow columns and plates. *Journal of Geophysical Research 110*, D17203 (August), 1–15.

Grenfell, T. C. 1983. A theoretical model of the optical properties of sea ice in the visible and near infrared. *Journal of Geophysical Research 88*, C14 (November), 9723–9735.

Grimaldi, F. M. 1665. *Physico-mathesis de lumine, coloribus, et iride.* Bononiæ. Reviewed in *Philosophical Transactions of the Royal Society of London*, Vol. 6, No. 79, pp. 3068–3070, 1671.

Groenhuis, R. A. J., Ferwerda, H. A., and Ten Bosch, J. J. 1983. Scattering and absorption of turbid materials determined from reflection measurements. 1: Theory. *Applied Optics 22*, 16 (August), 2456–2462.

Gutierrez, D., Munoz, A., Anson, O., and Seron, F. J. 2005. Nonlinear volume photon mapping. In *Rendering Techniques 2005 (Proceedings of Eurographics Symposium on Rendering)*, 291–300.

Haase, C. S., and Meyer, G. W. 1992. Modeling pigmented materials for realistic image synthesis. *ACM Transactions on Graphics 11*, 4 (October), 305–335.

Hale, G. M., and Querry, M. R. 1973. Optical constants of water in the 200-nm to 200-μm wavelength region. *Applied Optics 12*, 3 (March), 555–563.

Hall, R. A., and Greenberg, D. P. 1983. A testbed for realistic image synthesis. *IEEE Computer Graphics and Applications 3*, 8 (November), 10–20.

Hanrahan, P., and Krueger, W. 1993. Reflection from layered surfaces due to subsurface scattering. *Computer Graphics (Proceedings of ACM SIGGRAPH 93)* (August), 165–174.

Heisenberg, W. 1925. Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen. *Zeitschrift für Physik A: Hadrons and Nuclei 33*, 1 (December), 879–893.

Helmholtz, H. v. 1867. *Handbuch der Physiologischen Optik*. Leopold Voss. Three volumes published 1856–1866 and published together in *Algemeine Encyklopädie der Physik*, Vol. 9, 1867.

Henyey, L. G., and Greenstein, J. L. 1940. Diffuse radiation in the galaxy. *Annales d'Astrophysique 3*, 117–137. Also in *The Astrophysical Journal 93*, 1941.

Hero of Alexandria, ∼50 A.D., 1900. Catoptrics. In W. Schmidt, ed. and trans., *Heronis Alexandrini opera quae supersunt omnia*, Vol. 2:1, pp. 301–373, Teubner, 1900. English excerpts by Smith [1999].

Hertz, H. 1887. Ueber eine Einfluss des ultravioletten Lichtes auf die elektrische Entladung. *Annalen der Physik und Chemie 267*, 8, 983–1000.

Hertz, H. 1888. Ueber Strahlen elektrischer Kraft. *Sitzungsbericht der Berliner Akademie der Wissenschaften* (December). Reprinted in *Untersuchungen ueber die Ausbreitung der elektrischen Kraft*, pp. 184–198, Johann Ambrosius Barth, 1892.

Hobbes, T., 1644. Tractatus opticus. Book VII of the *Opticorum* in *Universæ geometriæ mixtæque mathematicæ synopsis, et bini refractionum demonstratarum tractatus*, Studio & Operâ F. M. Mersenni, pp. 567–589, 1644. English excerpts by Shapiro [1973].

Holliday, D. J., and Farin, G. E. 1999. A geometric interpretation of the diagonal of a tensor-product Bézier volume. *Computer Aided Geometric Design 16*, 8 (September), 837–840.

Holte, G. 1948. On a method of calculating the density of neutrons emitted from a point source in an infinite medium. *Arkiv för Matematik, Astronomi och Fysik 35A*, 1–9.

Hooke, R. 1665. *Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries thereupon*. Royal Society of London.

Huibers, P. D. T. 1997. Models for the wavelength dependence of the index of refraction of water. *Applied Optics 36*, 16 (June), 3785–3787.

Huygens, C. 1690. *Traité de la lumière: où sont expliquées les causes de ce qui luy arrive dans la réflexion, & dans la réfraction, et particuliérement dans l'étrange réfraction du cristal d'Islande.* Pierre vander Aa. Completed in 1678. Translated by S. P. Thompson, Dover Publications, Inc., 1962.

Ibn al-Haytham, ∼1016, 2001. Kitāb al-manāẓir. In A. M. Smith, *Alhacen's Theory of Visual Perception: A Critical Edition, with English Translation and Commentary, of the First Three Books of Alhacen's* De Aspectibus*, the Medieval Latin Version of Ibn al-Haytham's* Kitāb al-Manāẓir, Transactions of the American Philosophical Society, Vol. 91, parts 4–5, 2001.

Ibn Sahl, ∼984, 1993. On the burning instruments. In R. Rashed *Géométrie et dioptrique au X^e siècle: Ibn Sahl, al-Qūhī et Ibn al-Haytham*, Les Belles Lettres, 1993. English excerpts by Rashed [1990].

Ihrke, I., Ziegler, G., Tevs, A., Theobalt, C., Magnor, M., and Seidel, H.-P. 2007. Eikonal rendering: Efficient light transport in refractive objects. *ACM Transactions on Graphics 26*, 3 (July). Article 59.

Ishimaru, A. 1977. Theory and applications of wave propagation and scattering in random media. *Proceedings of the IEEE 65*, 7 (July), 1030–1061.

Ishimaru, A. 1978. *Wave Propagation in Random Media.* Academic Press, New York. Reissued by IEEE Press and Oxford University Press 1997.

Ishimaru, A. 1989. Diffusion of light in turbid material. *Applied Optics 28*, 12 (June), 2210–2215.

Jackèl, D., and Walter, B. 1997. Modeling and rendering of the atmosphere using Mie-scattering. *Computer Graphics Forum 16*, 4, 201–210.

James, R. W., and Firth, E. M. 1927. An X-ray study of the heat motions of the atoms in a rock-salt crystal. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 117*, 776 (December), 62–87.

Jenkins, M., and Falster, P. 1999. Array theory and Nial. Tech. rep., Electric Power Engineering Department, Technical University of Denmark, August.

Jenkins, M. A. 1981. A development system for testing array theory concepts. *ACM SIGAPL APL Quote Quad 12*, 1 (September), 152–159.

Jensen, H. W., Legakis, J., and Dorsey, J. 1999. Rendering of wet materials. In *Rendering Techniques '99 (Proceedings of the Tenth Eurographics Workshop on Rendering)*, Springer-Verlag, D. Lischinski and G. W. Larsoi, Eds., 273–282.

Jensen, H. W., Marschner, S. R., Levoy, M., and Hanrahan, P. 2001. A practical model for subsurface light transport. In *Proceedings of ACM SIGGRAPH 2001*, ACM Press, 511–518.

Johnson, T. E. 1963. *Sketchpad III, 3-D, Graphical Communication with a Digital Computer*. Master's thesis, MIT, Cambridge, Massachusetts.

Joule, J. P. 1843. On the calorific effect of magneto-electricity, and on the mechanical value of heat. *Report of the British Association for the Advancement of Science 12*.

Kajiya, J. T., and Von Herzen, B. P. 1984. Ray tracing volume densities. *Computer Graphics (Proceedings of ACM SIGGRAPH 84) 18*, 3 (July), 165–174.

Kajiya, J. T. 1985. Anisotropic reflection models. *Computer Graphics (Proceedings of ACM SIGGRAPH 85) 19*, 3 (July), 15–21.

Kajiya, J. T. 1986. The rendering equation. *Computer Graphics (Proceedings of ACM SIGGRAPH 86) 20*, 4 (August), 143–150.

Kattawar, G. W., and Plass, G. N. 1967. Electromagnetic scattering from absorbing spheres. *Applied Optics 6*, 8 (August), 1377–1382.

Kepler, J. 1604, 2000. *Optics: Paralipomena to Witelo and the Optical Part of Astronomy*. Green Lion Press. Translated by W. H. Donnahue from Kepler's *Ad Vitellionem paralipomena, quibus Astronomiæ pars optica traditur*, 1604.

Kepler, J. 1611. *Dioptrice seu Demonstratio eorum quæ visui & visibilibus propter conspicilla non ita pridem inventa accidunt*. Davidis Franci.

Kerker, M. 1969. *The Scattering of Light and Other Electromagnetic Radiation*. Academic Press, New York.

Kipfstuhl, J., Dieckmann, G., Oerter, H., Hellmer, H., and Graf, W. 1992. The origin of green icebergs in antarctica. *Journal of Geophysical Research 97*, C12 (December), 20,319–20,324.

Kirchhoff, G. 1860. Ueber das Verhältniß zwischen dem Emissionsvermögen und dem Absorptionsvermögen der Körper für Wärme und Licht. *Annalen der Physik und Chemie 185*, 2, 275–301.

Klassen, R. V. 1987. Modeling the effect of the atmosphere on light. *ACM Transactions on Graphics 6*, 3 (July), 215–237.

Kline, M., and Kay, I. W. 1965. *Electromagnetic Theory and Geometrical Optics*. Interscience Publishers (a division of John Wiley & Sons), New York.

Koziol, J. 1966. Studies on flavins in organic solvents I: Spectral characteristics of riboflavin, riboflavin tetrabutyrate and lumichrome. *Photochemistry and Photobiology 5*, 41–54.

kravtsov, Y. A. 1967. Complex rays and complex caustics. *Radiophysics and Quantum Electronics 10*, 9–10 (September), 719–730.

Krull, F. N. 1994. The origin of computer graphics within General Motors. *IEEE Annals of the History of Computing 16*, 3, 40–56.

Kuščer, I., and McCormick, N. J. 1991. Some analytical results for radiative transfer in thick atmospheres. *Transport Theory and Statistical Physics 20*, 351–381.

Kwan, A., Dudley, J., and Lantz, E. 2002. Who really discovered Snell's law? *Physics World 15*, 4 (April), 64.

Laërtius, D. ∼200 A.D., 1901. *The Lives and Opinions of Eminent Philosophers*. George Bell and Sons. Translated by C. D. Yonge.

Lambert, J. H. 1760. *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. Viduae Eberhardi Klett.

Laposky, B. F. 1953. *Oscillons: Electronic Abstractions*. Ben F. Laposky, Cherokee, Iowa.

Lee, Jr., R. L. 1990. Green icebergs and remote sensing. *Journal of the Optical Society of America A 7*, 10 (October), 1862–1874.

Lenard, P. 1902. Ueber die lichtelectrische Wirkung. *Annalen der Physik 313*, 149–197.

Leppäranta, M., and Manninen, T. 1988. The brine and gas content of sea ice with attention to low salinities and high temperatures. Tech. Rep. 2, Finnish Institute of Marine Research.

Leslie, J. 1804. *An Experimental Inquiry Into the Nature and Propagation of Heat*. printed for J. Mawman.

Levitt, T. 2000. Editing out caloric: Fresnel, Arago and the meaning of light. *The British Journal for the History of Science 33*, 1 (March), 49–65.

Lide, D. R., Ed. 2006. *CRC Handbook of Chemistry and Physics*, 87th ed. CRC Press.

Light, B., Eicken, H., Maykut, G. A., and Grenfell, T. C. 1998. The effect of included particulates on the spectral albedo of sea ice. *Journal of Geophysical Research 103*, C12 (November), 27,739–27,752.

Light, B., Maykut, G. A., and Grenfell, T. C. 2003. Effects of temperature on the microstructure of first-year arctic sea ice. *Journal of Geophysical Research 108*, C2, 3051 (February), **33**–1–16.

Light, B., Maykut, G. A., and Grenfell, T. C. 2004. A temperature-dependent, structural-optical model of first-year sea ice. *Journal of Geophysical Research 109*, C06013 (June), 1–16.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., and Garcia, H. E. 2006. Temperature. In *NOAA Atlas NESDIS 61*, S. Levitus, Ed., vol. 1 of *World Ocean Atlas 2005*. U.S. Government Printing Office, Washington, D.C. CD-ROM.

Lorentz, H. A. 1880. Ueber die Beziehung zwischen der Fortpflanzung des Lichtes und der Körperdichte. *Annalen der Physik und Chemie 245*, 4, 641–665.

Lorenz, L. 1880. Ueber die Refractionsconstante. *Annalen der Physik und Chemie 247*, 9, 70–103.

Lorenz, L. 1890. Lysbevægelser i og uden for en af plane Lysbølger belyst Kugle. *Det kongelig danske Videnskabernes Selskabs Skrifter*, 2–62. 6. Række, Naturvidenskabelig og Mathematisk Afdeling VI, 1.

Mackowski, D. W., Altenkirch, R. A., and Menguc, M. P. 1990. Internal absorption cross sections in a stratified sphere. *Applied Optics 29*, 10 (April), 1551–1559.

Malus, É. L. 1810. *Théorie de la double réfraction de la lumière dans les substances cristallines*. Gavnery.

Martin, P. A., and Rothen, F. 2004. *Many-Body Problems and Quantum Field Theory: An Introduction*, second ed. Springer.

Maxwell, J. C. 1873. *A Treatise on Electricity and Magnetism*. Clarendon Press. Two volumes.

Mayer, J. R. 1842. Bemerkungen über die Kräfte der unbelebten Natur. *Annalen der Chemie und Pharmacie 42*, 233–240.

Maykut, G. A., and Light, B. 1995. Refractive-index measurements in freezing sea-ice and sodium chloride brines. *Applied Optics 34*, 6 (February), 950–961.

Mehra, J., and Rechenberg, H. 1999. Planck's half-quanta: A history of the concept of zero-point energy. *Foundations of Physics 29*, 1, 91–132.

Meyer, G. W., and Greenberg, D. P. 1980. Perceptual color spaces for computer graphics. *Computer Graphics (Proceedings of ACM SIGGRAPH 80) 14*, 3 (July), 254–261.

Michalski, M.-C., Briard, V., and Michel, F. 2001. Optical parameters of milk fat globules for laser light scattering measurements. *Lait 81*, 787–796.

Middleton, W. E. K. 1964. The early history of the visibility problem. *Applied Optics 3*, 5, 599–602.

Mie, G. 1908. Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen. *Annalen der Physik 25*, 3, 377–445. IV. Folge.

Milonni, P. W., and Shih, M.-L. 1991. Zero-point energy in quantum theory. *American Journal of Physics 59*, 8, 684–698.

Milonni, P. W. 1994. *The Quantum Vacuum: An Introduction to Quantum Electrodynamics*. Academic Press.

Mishchenko, M. I., Travis, L. D., and Lacis, A. A. 2006. *Multiple Scattering of Light by Particles: Radiative Transfer and Coherent Backscattering*. Cambridge University Press.

Mishchenko, M. I. 2002. Vector radiative transfer equation for arbitrarily shaped and arbitrarily oriented particles: a microphysical derivation from statistical electromagnetics. *Applied Optics 41*, 33 (November), 7114–7134.

Mishchenko, M. I. 2003. Microphysical approach to polarized radiative transfer: extension to the case of an external observation point. *Applied Optics 42*, 24 (November), 4963–4967.

Moon, J. T., Walter, B., and Marschner, S. R. 2007. Rendering discrete random media using precomputed scattering solutions. In *Rendering Techniques 2007 (Proceedings of the 18th Eurographics Symposium on Rendering)*, 231–242+387.

Moravec, H. P. 1981. 3D graphics and the wave theory. *Computer Graphics (Proceedings of ACM SIGGRAPH 81) 15*, 3 (July), 254–261.

More, Jr, T. 1973. Axioms and theorems for a theory of arrays. *IBM Journal of research and development 17*, 2 (March), 135–175.

More, Jr, T. 1973. Notes on the development of a theory of arrays. Tech. Rep. 320-3016, IBM Philadelphia Scientific Center, Philadelphia, Pennsylvania, May.

More, Jr, T. 1973. Notes on the axioms for a theory of arrays. Tech. Rep. 320-3017, IBM Philadelphia Scientific Center, Philadelphia, Pennsylvania, May.

More, Jr, T. 1975. A theory of arrays with applications to databases. Tech. Rep. 320-2106, IBM Cambridge Scientific Center, Cambridge, Massachusetts, September.

More, Jr, T. 1976. Types and prototypes in a theory of arrays. Tech. Rep. 320-2112, IBM Cambridge Scientific Center, Cambridge, Massachusetts, May.

More, Jr, T. 1979. The nested rectangular array as a model of data. In *Proceedings of the International Conference on APL: Part 1*, 55–73.

More, Jr, T. 1981. Notes on the diagrams, logic and operations of array theory. In *Structures and Operations in Engineering and Management Systems*. TAPIR Publishers, Trondheim, Norway. Second Lerchendal book.

More, Jr, T. 1993. Transfinite nesting in array-theoretic figures, changes, rigs, and arms - Part I. *ACM SIGAPL APL Quote Quad 24*, 1 (August), 170–184.

Mossotti, O. F. 1850. Discussione analitica sullinfluenza che l'azione di un mezzo dielettrico ha sulla distribuzione dellelettricita alla superficie di piu corpi elettrici disseminati in esso. *Memorie di Matematica e de Fisica della Societa Italiana delle Science 24*, 2, 49–74.

Moulton, J. D. 1990. *Diffusion Modeling of Picosecond Laser Pulse Propagation in Turbid Media*. Master's thesis, McMaster University, Hamilton, Ontario, Canada.

Mulliken, R. S. 1925. The isotope effect in band spectra, II: The spectrum of boron monoxide. *Physical Review 25*, 3 (March), 259–294.

Mundy, W. C., Roux, J. A., and Smith, A. M. 1974. Mie scattering by spheres in an absorbing medium. *Journal of the Optical Society of America 64*, 12 (December), 1593–1597.

Møller, G. L. 1995. *On the Technology of Array-Based Logic*. PhD thesis, Electrical Power Engineering Department, Technical University of Denmark.

Narasimhan, S. G., Gupta, M., Donner, C., Ramamoorthi, R., Nayar, S. K., and Jensen, H. W. 2006. Acquiring scattering properties of participating media by dilution. *ACM Transactions on Graphics 25*, 3 (July), 1003–1012.

Neshyba, S. P., Grenfell, T. C., and Warren, S. G. 2003. Representation of a non-spherical ice particle by a collection of independent spheres for scattering and absorption of radiation: 2. Hexagonal columns and plates. *Journal of Geophysical Research 108*, D15, 4448 (August), **6**–1–18.

Newton, I. 1671. A letter of Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; containing his new theory about light and colors. *Philosophical Transactions of the Royal Society of London 6*, 80, 3075–3087.

Newton, I. 1704. *Opticks: or, a treatise of the reflexions, refractions, inflexions and colours of light*. Royal Society of London.

Nial. 2006. *The Nial Dictionary*, 6.3 ed. Nial Systems Limited, August.

Nicodemus, F. E., Richmond, J. C., Hsia, J. J., Ginsberg, I. W., and Limperis, T. 1977. Geometrical considerations and nomenclature for reflectance. Tech. rep., National Bureau of Standards (US), October.

Nishita, T., and Nakamae, E. 1983. Half-tone representation of 3-D objects illuminated by area sources or polyhedron sources. In *Proceedings of IEEE Computer Society's 7th International Computer Software & Applications Conference (COMPSAC 83)*, 237–242.

Nishita, T., and Nakamae, E. 1985. Continuous tone representation of 3-D objects taking account of shadows and interreflection. *Computer Graphics (Proceedings of ACM SIGGRAPH 85) 19*, 3 (July), 23–30.

Olson, D. W., White, C. H., and Richter, R. L. 2004. Effect of pressure and fat content on particle sizes in microfluidized milk. *Journal of Dairy Science 87*, 10, 3217–3223.

Palik, E. D., Ed. 1985. *Handbook of Optical Constants of Solids*. Academic Press.

Pattanaik, S. N., and Mudur, S. P. 1993. Computation of global illumination in a participating medium by Monte Carlo simulation. *The Journal of Visualization and Computer Animation 4*, 3, 133–152.

Pedersen, A., and Hansen, J. U. 1988. Q'Nial stand-by. Electric Power Engineering Department, Technical University of Denmark. Revised March 1993.

Pegau, W. S., Gray, D., and Zaneveld, J. R. V. 1997. Absorption and attenuation of visible and near-infrared light in water: Dependence on temperature and salinity. *Applied Optics 36*, 24 (August), 6035–6046.

Peirce, C. S. 1960. Grand logic (1893). In *Collected Papers of Charles Sanders Peirce: Elements of Logic*, C. Hartshone and P. Weiss, Eds., vol. II. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Perovich, D. K., and Govoni, J. W. 1991. Absorption coefficients of ice from 250 to 400 nm. *Geophysical Research Letters 18*, 7 (July), 1233–1235.

Pharr, M., and Hanrahan, P. 2000. Monte carlo evaluation of non-linear scattering equations for subsurface reflection. In *Proceedings of ACM SIGGRAPH 2000*, ACM Press, 75–84.

Pharr, M., and Humphreys, G. 2004. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann Publishers, an imprint of Elsevier Inc.

Phong, B. T. 1975. Illumination for computer-generated pictures. *Communications of the ACM 18*, 6 (June), 311–317.

Pierrat, R., Greffet, J.-J., and Carminati, R. 2006. Photon diffusion coefficient in scattering and absorbing media. *Journal of the Optical Society of America A 23*, 5 (May), 1106–1110.

Planck, M. 1900. Ueber eine Verbesserung der Wien'schen Spectralgleichung. *Verhandlungen der Deutschen Physikalischen Gesellschaft 2*, 13, 202–204.

Planck, M. 1900. Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft 2*, 17, 237–245.

Planck, M. 1912. Über die Begründung des Gesetzes der scwarzen Strahlung. *Annalen der Physik 342*, 642–656.

Planck, M. 1914. *The Theory of Heat Radiation*. F. Blaiston Son & Co. English translation of the second German edition (*Vorlesungen über dir Theorie der Wärmestrahlung*) reprinted by Dover Publications, Inc., in 1991.

Plato, ∼360 B.C., 1989. Timaeus. Translated by B. Jowett in *The Collected Dialogues of Plato*, E. Hamilton and H. Cairns ed., Bollinger Series 71, Princeton University Press, 1989.

Pope, R. M., and Fry, E. S. 1997. Absorption spectrum (380–700 nm) of pure water. II. integrating cavity measurements. *Applied Optics 36*, 33 (November), 8710–8723.

Poynting, J. H. 1884. On the transfer of energy in the electromagnetic field. *Philosophical Transactions of the Royal Society of London 175*, 343–361.

Preisendorfer, R. W. 1965. *Radiative Transfer on Discrete Spaces*. Pergamon Press.

Ptolemaeus, C., ∼160 A.D., 1996. Optics. Translated by A. M. Smith in *Ptolemy's Theory of Visual Perception: An English Translation of the* Optics *with Introduction and Commentary*, Transactions of the American Philosophical Society, Vol. 86.2, 1996.

Quan, X., and Fry, E. S. 1995. Empirical equation for the index of refraction of seawater. *Applied Optics 34*, 18 (June), 3477–3480.

Randrianalisoa, J., Baillis, D., and Pilon, L. 2006. Modeling radiation characteristics of semitransparent media containing bubbles or particles. *Journal of the Optical Society of America A 23*, 7 (July), 1645–1656.

Rapaport, D. C. 2004. *The Art of Molecular Dynamics Simulation*, second ed. Cambridge University Press.

Rashed, R. 1990. A pioneer in anaclastics: Ibn Sahl on burning mirrors and lenses. *Isis 81*, 464–491.

Richardson, C. 1976. Phase relationships in sea ice as a function of temperature. *Journal of Glaciology 17*, 77, 507–519.

Ripoll, J., Yessayan, D., Zacharakis, G., and Ntziachristos, V. 2005. Experimental determination of photon propagation in highly absorbing and scattering media. *Journal of the Optical Society of America A 22*, 3 (March), 546–551.

Roberts, L. G. 1963. *Machine Perception of Three-Dimensional Solids*. PhD thesis, MIT, Cambridge, Massachusetts.

Rømer, O. 1676. Démonstration touchant le mouvement de la lumière trouvé. *Journal des Sçavans* (December), 276–279. Translated in *Philosophical Transactions of the Royal Society of London*, Vol. 12, No. 136, pp. 893–894, 1677–1678.

Rubens, H., and Kurlbaum, F. 1900. Über die Emission langwelliger Wärmestrahlen durch den schwarzen Körper bei verschiedenen Temperaturen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin 33*, 2, 929–941.

Rushmeier, H. 1995. Input for participating media. In *Realistic Input for Realistic Images*, ACM SIGGRAPH 95 Course Notes, ACM Press. Also appeared in the ACM SIGGRAPH 98 Course Notes - A Basic Guide to Global Illumination.

Saleh, B. E. A., and Teich, M. C. 2007. *Fundamentals of Photonics*, second ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Schiebener, P., Straub, J., Sengers, J. M. H. L., and Gallagher, J. S. 1990. Refractive index of water and steam as function of wavelength, temperature and density. *Journal of Physical and Chemical Reference Data 19*, 3, 677–717.

Schmidt, D. G., Walstra, P., and Buchheim, W. 1973. The size distribution of casein micelles in cow's milk. *Netherland's Milk Dairy Journal 27*, 128–142.

Schrödinger, E. 1926. An undulatory theory of the mechanics of atoms and molecules. *Physical Review 28*, 6 (December), 1049–1070.

Schuster, A. 1905. Radiation through a foggy atmosphere. *The Astrophysical Journal 21*, 1 (January), 1–22.

Schwarzschild, K. 1906. Ueber das Gleichgewicht der Sonnenatmosphäre. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse 1906*, 1, 41–53.

Shapiro, A. E. 1973. Kinematic optics: A study of the wave theory of light in the seventeenth century. *Archive for History of Exact Sciences 11*, 2–3 (June), 134–266.

Sharma, G. 2002. Comparative evaluation of color characterization and gamut of LCDs and CRTs. In *Proceedings of SPIE, Color Imaging: Device-Independent Color, Color Hardcopy, and Applications VII*, R. Eschbach and G. Marcu, Eds., vol. 4663, 177–186.

Smith, A. M. 1999. *Ptolemy and the Foundations of Ancient Mathematical Optics: A Source Based Guided Study*, vol. 89.3 of *Transactions of the American Philosophical Society*. American Philosophical Society, Philadelphia.

Smits, B. E., and Meyer, G. W. 1990. Newton's colors: Simulating interference phenomena in realistic image synthesis. In *Proceedings of Eurographics Workshop on Photosimulation, Realism, and Physics in Computer Graphics*, 185–194.

Sommerfeld, A., and Runge, J. 1911. Anwendung der Vektorrechnung auf die Grundlagen der Geometrischen Optik. *Annalen der Physik 340*, 277–298.

Stam, J., and Fiume, E. 1995. Depicting fire and other gaseous phenomena using diffusion processes. In *Proceedings of ACM SIGGRAPH 1995*, ACM Press, 129–136.

Stam, J., and Languénou, E. 1996. Ray tracing in non-constant media. In *Rendering Techniques '96 (Proceedings of the Seventh Eurographics Workshop on Rendering)*, X. Pueyo and P. Schröder, Eds., 225–234.

Stam, J. 1995. Multiple scattering as a diffusion process. In *Rendering Techniques '95 (Proceedings of the Sixth Eurographics Workshop on Rendering)*, Springer-Verlag, P. Hanrahan and W. Purgathofer, Eds., 41–50.

Stam, J. 1999. Diffraction shaders. In *Proceedings of ACM SIGGRAPH 1999*, ACM Press, 101–110.

Stefan, J. 1879. Über die Beziehung zwischen der Wärmestrahlung und der Temperatur. *Sitzungsberichte der Matematischen-Naturwissenschaftlichen Classe der Kaiserlichen Akademie der Wissenschaften 79*, 2, 391–428.

Stelson, A. W. 1990. Urban aerosol refractive index prediction by partial molar refraction approach. *Environmental Science & Technology 24*, 11, 1676–1679.

Stiles, W. S., and Burche, J. M. 1959. N.P.L. colour-matching investigation: Final report (1958). *Optica Acta 6*, 1–26.

Stockman, A., and Sharpe, L. T. 2000. The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research 40*, 13, 1711–1737.

Strutt, J. W. 1871. On the light from the sky, its polarization and colour. *Philosophical Magazine 41*, 107–120, 274–279. Reprinted in *Scientific Papers by John William Strutt, Baron Rayleigh*, Vol. 1, No. 8, pp. 87–103, Cambridge University Press, 1899.

Sudarshan, E. C. G. 1981. Quantum theory of radiative transfer. *Physical Review A 23*, 6 (June), 2802–2809.

Sutherland, I. E. 1963. *Sketchpad. A Man-Machine Graphical Communication System*. PhD thesis, MIT, Cambridge, Massachusetts.

Tannenbaum, D. C., Tannenbaum, P., and Wozny, M. J. 1994. Polarization and birefringency considerations in rendering. In *Proceedings of ACM SIGGRAPH 1994*, ACM Press, 221–222.

Theophrastus, ∼300 B.C., 1917. On the Senses. Translated by G. M. Stratton in *Theophrastus and the Greek Physiological Psychology before Aristotle*, Allen and Unwin, 1917.

Thomas, S. W. 1986. Dispersive refraction in ray tracing. *The Visual Computer 2*, 1 (January), 3–8.

Thompson, T. G., and Nelson, K. H. 1956. Concentration of brines and deposition of salts from sea water under frigid conditions. *American Journal of Science 254*, 4 (April), 227.

Tong, X., Wang, J., Lin, S., Guo, B., and Shum, H. 2005. Modeling and rendering of quasi-homogeneous materials. *ACM Transactions on Graphics 2005 24*, 3, 1054–1061.

van de Hulst, H. C. 1949. On the attenuation of plane waves by obstacles of arbitrary size and form. *Physica 15*, 8–9 (September), 740–746.

van de Hulst, H. C. 1957. *Light Scattering by Small Particles*. John Wiley & Sons, Inc., New York. Unabridged and corrected version of the work published by Dover Publications, Inc., in 1981.

von Neumann, J. 1955. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press. Translated from the German edition by Robert T. Beyer.

Waller, I. 1946. On the theory of the diffusion and the slowing down of neutrons. *Arkiv för Matematik, Astronomi och Fysik 34A*, 3, 1–9.

Walstra, P., and Jenness, R. 1984. *Dairy Chemistry and Physics*. John Wiley & Sons, New York.

Walstra, P. 1975. Effect of homogenization on the fat globule size distribution in milk. *Netherland's Milk Dairy Journal 29*, 279–294.

Warnock, J. E. 1969. A hidden surface algorithm for computer generated halftone pictures. Tech. Rep. 4-15, University of Utah, June.

Warren, S. G., Roesler, C. S., Morgan, V. I., Brandt, R. E., Goodwin, I. D., and Allison, I. 1993. Green icebergs formed by freezing of organic-rich seawater to the base of antarctic ice shelves. *Journal of Geophysical Research 98*, C4 (April), 6921–6928.

Warren, S. G., Brandt, R. E., and Grenfell, T. C. 2006. Visible and near-ultraviolet absorption spectrum of ice from transmission of solar radiation into snow. *Applied Optics 45*, 21 (July), 5320–5334.

Warren, S. G. 1984. Optical constants of ice from the ultraviolet to the microwave. *Applied Optics 23*, 8 (April), 1206–1225.

Weinberg, A. M., and Wigner, E. P. 1958. *The Physical Theory of Neutron Chain Reactors*. The University of Chicago Press, Chicago, Illinois.

Wen, B., Tsang, L., Winebrenner, D. P., and Ishimaru, A. 1990. Dense medium radiative transfer: Comparison with experiment and application to microwave remote sensing and polarimetry. *IEEE Transactions on Geoscience and Remote Sensing 28*, 1 (January), 46–59.

Whitted, T. 1980. An improved illumination model for shaded display. *Communications of the ACM 23*, 6 (June), 343–349.

Wien, W. 1896. Ueber die Energievertheilung im Emissionsspectrum eines schwarzen Körpers. *Annalen der Physik und Chemie 294*, 662–669.

Wiscombe, W. J. 1980. Improved Mie scattering algorithms. *Applied Optics 19*, 9 (May), 1505–1509.

Wolf, E. 1976. New theory of radiative energy transfer in free electromagnetic fields. *Physical Review D 13*, 4 (February), 869–886.

Wollaston, W. H. 1802. On the double refraction of Iceland crystal. *Philosophical Transactions of the Royal Society of London 92*, 381–386.

Wu, Z. S., and Wang, Y. P. 1991. Electromagnetic scattering for multilayered sphere: Recursive algorithms. *Radio Science 26*, 6, 1393–1401.

Yang, P., Gao, B.-C., Wiscombe, W. J., Mischenko, M. I., Platnick, S. E., Huang, H.-L., Baum, B. A., Hu, Y. X., Winker, D. M., Tsay, S.-C., and Park, S. K. 2002. Inherent and apparent scattering properties of coated or uncoated spheres embedded in an absorbing host medium. *Applied Optics 41*, 15 (May), 2740–2758.

Yang, W. 2003. Improved recursive algorithm for light scattering by a multilayered sphere. *Applied Optics 42*, 9 (March), 1710–1720.

Yin, J., and Pilon, L. 2006. Efficiency factors and radiation characteristics of spherical scatterers in an absorbing medium. *Journal of the Optical Society of America A 23*, 11 (November), 2784–2796.

Young, T. 1802. An account of some cases of the production of colours, not hitherto described. *Philosophical Transactions of the Royal Society of London 92*, 387–397.

Index